Digital Citizen Engagement
A Case Study Approach Involving Multiple Countries and Platforms

Agarwal, Pushkal

Awarding institution:
King's College London

Digital Citizen Engagement: A Case Study Approach Involving Multiple Countries and Platforms

Pushkal Agarwal

Department of Engineering, Faculty of Natural and Mathematical Sciences

King's College London

This dissertation is submitted for the degree of

Doctor of Philosophy

July 2022

ABSTRACT

Disengagement and disenchantment with the Parliamentary process is an important concern in today's Western democracies. I distill the ways in which three distinct kinds of actors: politicians, citizens, digital mass media platforms (news websites and social platforms), together with a facilitator - the Internet around the world are therefore seeking new ways to engage with each other. Generalising these online interactions as a three-actors model of online political communications, in this dissertation I collect large datasets from multiple digital platforms in multiple countries and seek to **understand digital citizen engagement as political communication in the modern context**. Using large datasets I develop methods and metrics to study various interactions between different nodes in the model, namely in three parts: politicians with citizens, politicians with digital mass media platforms and citizens with digital mass media platforms.

In the first part, I study the interactions between politicians and citizens on Twitter in the United Kingdom (UK). I collect a large dataset of 2.5 Million tweets (involving all 579 out of 650 UK MPs who are active on Twitter) and seek to shed light on political communication by examining the volume and nature of the interaction between MPs and citizens. I find that Twitter is a venue with substantial volume of dialogue and cross-party interaction between MPs of one party and citizens who support (follow) MPs of other parties. Increasing online activity in the political and other spheres has led to concerns about online hate. By combining widely-used hate speech detection tools trained on several widely available datasets, I identify hate in 2.5 Million tweets against MPs and find that approximately 1% of the total volume of tweets corresponds to known volumes of hate speech on Twitter. I find that MPs are subject to intense 'pile on' hate by citizens whereby they get more hate when they are already busy with a high volume of mentions.

In the second part, I study interactions between politicians (including their supporters) with digital mass media platforms which are used for content propagation largely around politics (or politicians) and political ideologies. I study this interaction in two sub-parts: news media platforms exhibiting political ideologies of a party and public profiles of politicians exhibiting ideologies of political actors. Though polarisation has been extensively studied

with respect to content, it is still unknown how it associates with the online tracking experienced by users, especially when they exhibit certain demographic characteristics. The first sub-part includes interactions with hyper-partisan news websites (having self-announced political ideologies), namely left or right-leaning. I design and deploy a methodology to systematically measure differences in user tracking based on user personas (left or right). I collect 3.5 Million cookies while testing 9 personas on 556 hyper-partisan websites. I find that right-leaning websites tend to track users more intensely than left-leaning and partisan news websites change their tracking behaviour, depending on the demographic of the user. In the second sub-part, I also offer a complementary study which investigates the presence of polarisation in Wikipedia profiles of politicians using 650 UK MP pages edits (from 2002 to 2019). I collect in total 231k edits, which are made by 43k unique editors. I find that most editors specialize in a specific party and choose specific news outlets as references.

Finally in the third part, I characterise interactions between citizens with digital mass media platforms. I deep dive into the use of platforms by citizens in emerging economies by characterising a multilingual social network in India. I work on an extensive dataset consisting of over 1.2 million posts across 14 languages posted on a platform called ShareChat. I augment this data in multiple ways which also include large-scale clustering of semantically similar posts. I find that certain regional languages tend to be dominant in soliciting political memes and images. Due the varying population sizes, I see a strong skew towards widely spoken languages. Surprisingly, Hindi is not the most popular language though. Instead, Telugu and Malayalam accumulate the majority of posts (16.4% and 15% respectively). However, I find that posts from Hindi (and images having text in English) have the largest cross-lingual diffusion. With new users using such platforms, cases of involvement in scams are also increasing. I investigate this challenge and present a case-study on the abuse of joining links of 5,051 public WhatsApp groups where a small set of users are sending unwanted junk messages. I cluster these junk messages and characterise the junk content shared within public groups. I find that nearly 1 in 10 messages is an unwanted junk message.

In a nutshell, in this dissertation I contribute models, metrics and datasets which quantify engagement in political communication across existing and new emerging digital platforms for political discourse.

– To Mom, Dad, my family and Lord Krishna.

ACKNOWLEDGEMENTS

This dissertation is not just a destination, rather an exciting journey with life-long learning. First and foremost, I would like to thank my supervisors Nishanth Sastry and Edward Wood for their continuous guidance, motivation and support during all times. I specially thank Nishanth for setting high expectations and at the same time providing key motivation to meet those expectations. I will never forget that one line of motivation he said during our conversations – *There is no last hope*. Also, this dissertation would not have been possible without the unconditional support from Edward from the House of Commons. Under his guidance, for the first time I learned components of political communication and the UK Parliament. These helped me in developing a career which involves inter-disciplinary research. I also thank my other co-supervisors Toktam Mahmoodi and Andrew Blick for their inputs during different stages of my PhD.

I am lucky to have amazing teammates from King's College London who helped and guided me in challenging times. I would like to thank my seniors Aravindh and Sagar for their key support and unlimited fun memories, like our table-tennis sessions. I also thank Changtao, Dmytro, Emeka, Tooba, Rachel, Abdul, Vibhor, Shounak and Margarita for providing me time to time feedback on my research, sometimes in return for a free coffee. As this PhD is a journey to me, I am fortunate that crossed paths with many great researchers. I would like to thank my co-authors and mentors Nicolas Kourtellis (Telefonica), Miriam Redi (Wikimedia), Luca Maria Aiello (Nokia Bell Labs), Gareth Tyson (Queen Mary University), Kiran Garimella (MIT) and Sakthi Balan (LNMIIT). I am grateful to Sakthi Balan with whom I wrote my first research paper during my bachelors as well as a research paper during the final year of my PhD.

A big shout-out to all my friends Aishwarya, Akash, Dikshita, Divyansh, Monica, Navneet, Mukul, Shailesh and Shikhar who were there to encourage me through calls and meet-ups. Finally, I would like to thank my Mom, Dad and family for their unconditional love. I thank my brother (Vidit) who always inspired me by telling his stories when he was in my shoes, i.e. his PhD. Last but not the least, I sincerely thank Lord Krishna, the one who's teachings from *Shri Bhagwat Geetha* have helped me at various integral phases of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

We are all now connected by the Internet, like neurons in a giant brain.

Professor Stephen Hawking

1.1 Introduction to Digital Citizen Engagement

In 21st century the political discourse is not only driven by state or authorised news but also a large fraction now engages with multiple different platforms. Political communication is defined by Brian McNair as "purposeful communication about politics" [McN11]. It includes *(inter alia)* communication about politicians and other political actors, and their activities, as contained in news reports, editorials, and other forms of media discussion of politics such as social media and large-scale online encyclopedias like Wikipedia. These communications are not just limited to textual speeches. Rather, they include a wide range of presentation with visuals like logos, dress, memes (visuals with brief text) and much more. Clearly, with the recent advancements in the use of digital platforms the classic nature of political communications, i.e. 'broadcasting' into media has now seen a shift to more 'direct' representation to new online media which makes it much easier to put information into the public domain and reach potentially millions of people [Col05a, Col05a]. One such example is Twitter, a micro-blogging website that Brants & Voltmer study to understand

the shift from mediatisation to de-centralisation of political communication in postmodern democracy [BV11].

In this dissertation, I use large datasets and timely case studies around social media platforms to understand digital citizen engagement in contemporary political communication using a model as proposed by Balčytienė & Vinciūnienė [BV08]. Figure 1.1 shows this model with three actors: *Politicians, Citizens and Digital Mass Media Platforms* along with a factor the *Internet*, facilitating the rest. I aim to understand each node and it's interaction with others as shown in Figure 1.1. The model shows bi-directional digital communication between Politicians (P: involving political actors who are mostly elected by members of the public), Mass Media (M: involving reports on social media, news media, online groups and other digital platforms) and Citizens (C: who interact with politicians, media platforms or with other citizens on digital platforms). This political communication exists in digital ecosystem which is connected through Internet (I).

Early studies such as McNair's model [McN11] does not connect Politicians and Citizens directly, rather the communication was thought to be carried out using mass media. Given recent changes in political communication as studied by Brants & Voltmer [BV11], Figure 1.1 shows that there exist a bi-directional online engagement between Politicians and Citizens as well. This is also in-line with Coleman's early discussion of emergence of direct representation using digital media [Col05b]. For example– taking the case of digital platforms such as Twitter in the UK, the dataset I collect as part of this research shows that in recent years nearly all (579 out of 650) UK Member of the Parliament in House of Commons (generally abbreviated as MPs) who are elected by citizens have created Twitter accounts, and have amassed huge followings comparable to a sizable fraction of the country's population. Moreover, I also find that MPs manage their interaction strategically, replying selectively to UK based citizens and thereby serving in their role as elected representatives, and using retweets to spread their party's message.

Next, I provide an overview of interaction between each pair of nodes shown in Figure 1.1 with the facilitation of Internet. In this I also discuss new challenges in such political communication and their underlying research question(s) which I study in later chapters

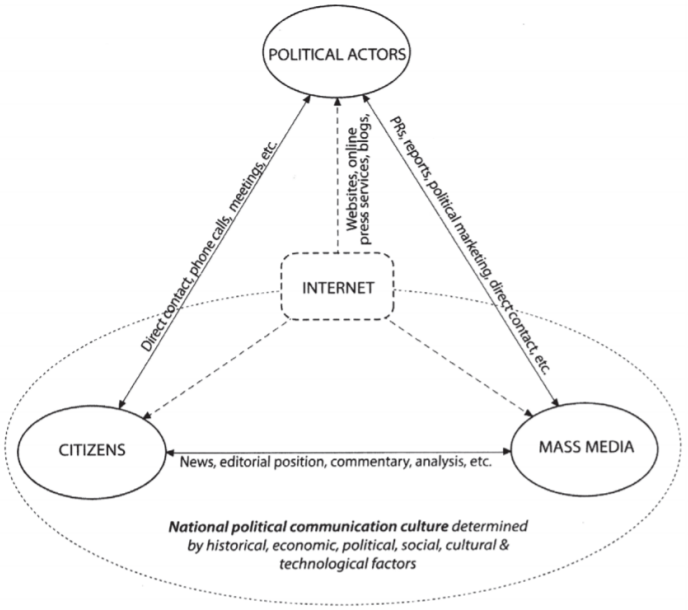


Fig. 1.1 Politicians, citizens and digital mass media platforms in contemporary political communication [BV08] which is connected through the Internet.

of dissertation. Following this, there are list of my contributions, structural overview of dissertation and finally list of peer reviewed papers which are part of my research.

1.2 Interaction between Politicians and Citizens

There has been much bemoaning the apparent decrease in political engagement amongst the electorate in western democracies such as the UK [Hou17]. Connecting elected representatives such as Members of Parliament (MPs)[1] with voters (especially young voters) is seen as a means to "revive democracy" [Rt 02] and recently there is much hope that online methods such as Twitter will play a key role in this [Spe15]. Yet, traditional scholarship on legislative studies has focused mostly on the relationship between the Parliament and the Government, casting MPs in the core roles of legislation and scrutiny of the Executive branch, neglecting the communication between citizens and their MPs [LB12].

[1] I interchangeably use the terms "Politicians" and "MPs" to refer elected members in the Parliament

To frame the discussion, I turn to the nature of the discourse between citizens and UK MPs on Twitter where a remarkable 579 out of 650 Members of Parliament (MPs) were active in 2017. This represents a dramatic rise from just a few years back: only in 2011, there were just 51 MPs "dipping their toes" in Twitter [JL11]. These MPs have a collective following of 12.83, 2.5 Million incoming tweets from citizens and 176k post by them during the study period (Oct 2017–Nov 2017). I consider attention load and its management as two-way communication happening between MPs and citizens. I also discuss the nearest offline equivalent for interaction between MPs and citizens – *constituency service*. As a new and additional medium of citizen engagement, I propose following research question around this large Twitter dataset:

RQ-1: How much load does social media place on MPs, and how does this load vary? Do MPs selectively prioritise and reply to certain forms of engagement? What is the nature and tone of the conversations?

I discuss this research question in Chapter 3. I find that although there is a huge amount of information overload during short periods, MPs strategically manage their relationship with their followers and this information overload by balancing their different roles as representatives of their constituency and their party [Sta08]: They selectively reply to user profiles in the UK and within their constituency region, fulfilling their representative role, and use retweets as a mechanism to promote their image and spread the message of their party.

For a deeper understanding of positive and negative impact of these interactions I then present two case studies in Chapter 3. These case studies are again based on the above research questions and collected dataset. The first case study discusses the novel use of Twitter by MP Chris Bryant to take citizens' poll. In this way citizens' also took part in proposing a Bill in Parliament. With the increase in usage of internet, its misuse has also increased [FW20, GGR$^+$18, BV11]. Hence, in the second case study, I discuss concerns

about 'online hate' as a growing problem. Using the significant cross-party interaction, between citizens who support and follow MPs from one party, and MPs from other parties I find that within each party, the ratio of online hate is less and more hate is coming from across parties. Hate users also target to some specific demographics of MPs.

Ethics note: The study and dataset has been registered with Kings Data Protection[2]. The public dataset used to study the research question of this section is made available at https://bit.ly/2HSTOsa for research usage. Following Twitter's Terms of Service[3], only the Tweet IDs are shared.

1.3 Interaction between Politicians and Digital Mass Media Platforms

With most of the political discussion happening in the cyber-space, political parties and publishers now involve more in providing user audience their news content via online means like websites, social media and Wikipedia (an online encyclopedia). Contrary to the traditional news media (e.g., newspapers, magazines, TV), on the web, users tend to receive personalised content tailored to their interests. There is an abundance of companies helping publishers to serve dynamically customised content and targeted ads to visitors based on their characteristics and behaviour to support conversion and long-term engagement goals. Another selective political initiative is the highly polarized Hyper-partisan news websites (HPWs) in US politics, studied by [BJBS18]. HPWs declare themselves right- or left-leaning in their web portals. Interestingly, many of these HPWs were deleted after US 2016 elections. However, more than 550 still exist and I focus my study on them. I study the interaction between Politicians and Digital Mass Media Platforms through the lens of these online websites where users consume information around political news, often along political ideologies. At such a juncture, I ask following research question to study such interactions:

[2]https://kdpr.kcl.ac.uk/Personalisation/DisplayPage/8
[3]https://twitter.com/en/tos

RQ-2: Do these websites – which have been shown to have highly selective audiences – exhibit any particular behaviour when it comes to tracking and content delivery to their online visitors?

I answer this multi-faceted question in Chapter 4, along with two case studies. I first establish a methodology for understanding how HPWs and their embedded third parties track different user demographics. I create 9 carefully crafted personas representing different genders and age groups. I load browsers with these personas and visit a list of 556 verified HPWs [BJBS18], to observe differences in the way these personas are being tracked via different types of cookies placed by HPWs and the third party ad-ecosystem. I make use of *OpenWPM* [EN20], a popular tool for measurements and automating web browsers. Using OpenWPM, over multiple crawls (during Nov 2018–Dec 2018) I collect nearly 3.5 Million cookies and 20 Million collective HTTP requests, responses and redirects. I find that in HPWs , advertisers set the majority of cookies on users' browsers. These advertisers are among the top in the overall ad-ecosystem, and are over-represented on HPWs compared to their presence on the general web. Also, having an established persona from a particular demographic (with cookies obtained from visiting stereotypical websites for users of that demographic) results in *up to 15% more cookies* stored than for a baseline with no set persona. Furthermore, popular or highly-ranked HPWs track users more intensely than lower-ranked websites. More importantly, my findings show that right-leaning websites, in general, track users with *up to 25% more cookies* than left-leaning websites. Additionally, I discuss implications of my work for to understand new emerging online issue *i.e.,* polarisation. I present this in two case studies in Chapter 4. In the first case study, I search for large-scale patterns to examine the extent to which the ad-ecosystem performs unbalanced tracking. I achieve this by co-clustering both the personas and websites visited, using Non-Negative Matrix (NMF) factorization [LS99, LGG18, ZYCG07]. Unfortunately, this growth of online political communications has been accompanied by rising partisanship and hence I see that

certain trackers are predominant in a preferentially attached persona exhibiting a certain political leaning. In the second case study, I find that for readers who go to Wikipedia a lot is known about the general usage and information consumption but, less is known about the life-cycle and neutrality of Wikipedia articles in the context of politics. I find that there exist huge peaks of attention during election times, related to signs of engagement on other social media (e.g. Twitter). I also quantify editors' polarisation and find that most editors specialise in a specific party and choose specific news outlets as references.

Ethics note: To study the research question in this section, I extend an existing crawling tool with my methodology, and open source the framework[4] in order to enable fellow researchers, policy makers or even end-users to audit websites on how they personalize tracking technology based on the visitor's web profile. The dataset does not involve any real users or any personal data. The dataset I collect in the second case study is also available for the research community[5]. The study and dataset has been registered with Kings Data Protection[6].

1.4 Interaction between Citizens and Digital Mass Media Platforms

I study these interactions of citizens and digital media platforms using datasets of India, specifically around understudied national elections of world's largest democracy. Given the rich linguistic diversity in India, I wish to understand the manner in which such a large scale *national* effort plays out on social media despite the differences in languages among the citizen of India. It is undeniable that social media played an important role in the Indian General Elections, helping mobilise voters and spreading information about the different major parties [Rao19, Pat19a, ZCDC+19, McN11]. I focus my efforts on understanding whether and to what extent information on social media users crosses regional and linguistic divides. To explore this, I formulate the following research question:

[4] Dataandcodeavailablefromhttps://nms.kcl.ac.uk/netsys/datasets/partisan-tracking/
[5] tiny.cc/wikipedia-mps
[6] https://kdpr.kcl.ac.uk/Personalisation/DisplayPage/8

RQ-3: Can some content transcend language silos and how often does this happen? Amongst which languages and do their semantic meanings mutate? What are the new emerging challenges which platforms need to consider for tackling misbehaviour?

To explore this in Indian context, I study a large-scale dataset, consisting of over 1.2 million posts across 14 languages during and before the election campaigning period, from **ShareChat– Made in India Multi-lingual Social Media Platform** in Chapter 5. This is a media and text-sharing platform designed for Indian users, with over 50 million monthly active users in India. It has one major difference that aids my research design: Unlike other global platforms, different languages are separated into communities of their own, which creates a silo effect and a strong sense of commonality. Despite this regional appeal, it has accumulated tens of millions of users in just 4 years of existence, with most being first time Internet users [Lev19].

Building on these users sharing and engaging patterns across languages I present answers to second and third part of the aforementioned research question in two case studies. In the first case study (c.f. §5.5), after hand-coding and manual translations of posts I find that for some languages and posts their semantic meanings mutate. I find cases where same images (political memes) are altered to make them more consumable in a specific language, or the meaning is altered on purpose in the shared text. In the second case (c.f. §5.6), I characterise junk messaging on an emerging and understudied platform, WhatsApp. With over 1.5 billion global active users each day and over 5 billion worldwide downloads from the Android Play Store alone, WhatsApp has become a key part of the communications landscape. I characterise junk content shared within public groups and find that nearly 1 in 10 messages is junk around different topics, languages and users. My results have key implications for understanding political communications in multi-lingual environments. Next, I list contribution and provide an overview of the structure of this dissertation.

Ethics note: The shared datasets I use to study this research question abides by the terms of service of WhatsApp and Sharechat. These data collections were considered exempt by the Institutional Review Board at MIT. All data was anonymised before analysis, and any personally identifiable information was masked. An anonymised version of the dataset is available for noncommercial research usage from https://tiny.cc/share-chat and http://tiny.cc/netsys-whats-app.

1.5 Contributions and Overview

1.5.1 Contributions

Contributions of this dissertation are in three fold. First, I collect large-scale datasets around online political communication and make them openly available for research use. Wherever applicable I make processed and anonymised version of data available, following the platform's terms of services with open access to peer-reviewed papers[7]. Secondly, my characterisation and findings which are included in this dissertation are helpful in understating the new forms of engagement. Recent media articles include Wired News[8], King's College Research Updates[9], University of Surrey News[10] and a citation by Parliament of South Australia[11]. Third contribution includes generalised models and methodology which can be replicated to understand other digital platforms in multiple other countries. I cover operationalisation of these methodologies as example case studies. These contributed models and methodologies are in use by other researchers who have published works in collaboration (c.f. §1.5.3).

[7] Datasets from this research are hosted at nms.kcl.ac.uk/netsys/datasets/
[8] https://www.wired.com/story/right-left-news-site-ad-tracking/
[9] https://www.kcl.ac.uk/partisan-us-news-websites-track-user-data-more-than-general-web
[10] https://www.surrey.ac.uk/news/hate-speech-detection-tools-show-greater-online-abuse-mps-bame-backgrounds
[11] Joshua Forkert and Parliamentary Officer. "Guidelines on the use of social media in Parliament."

1.5.2 List of Peer Reviewed Publications

Following is a list of publications which are led by author as a part of this research work and discuss in later chapters. The next list (c.f. §1.5.3) is of publications which are extensions of this research in collaboration[12].

Chapter 3:

- **Agarwal P**, Sastry N, Wood E. 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 (ICWSM) 2019 Jul 6 (Vol. 13, pp. 26-37). [ASW19] [🎗 Best research impact award at the Poster competition of King's College London.]

- **Agarwal P**, Hawkins O, Amaxopoulou M, Dempsey N, Sastry N, Wood E. Hate Speech in Political Discourse: A Case Study of UK MPs on Twitter. In Proceedings of the ACM Conference on Hypertext and Social Media (HT) 2021 Aug 25 (pp. 5-16). [AHA$^+$21] [Best poster presentation at the Study of Parliament Group Annual Weekend (2022)]

Chapter 4:

- **Agarwal P**, Joglekar S, Papadopoulos P, Sastry N, Kourtellis N. Stop Tracking me Bro! Differential Tracking of User Demographics on Hyper-partisan Websites. In Proceedings of The Web Conference (WWW) 2020 Apr 20 (pp. 1479-1490). [AJP$^+$20]

- **Agarwal P**, Redi M, Sastry N, Wood E, Blick A. Wikipedia and Westminster: Quality and Dynamics of Wikipedia Pages about UK Politicians. In Proceedings of the ACM Conference on Hypertext and Social Media (HT) 2020 Jul 13 (pp. 161-166). [ARS$^+$20]

Chapter 5:

- **Agarwal P**, Garimella K, Joglekar S, Sastry N, Tyson G. Characterising User Content on a Multi-lingual Social Network. In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2020 May 26 (Vol. 14, pp. 2-11). [AGJ$^+$20]

[12]These works are not main contributions of the author, hence they are not included in this dissertation. Author collaborated and provided guidance to the team members which helped them in extending author's main contributed publications.

- **Agarwal P**, Raman A, Ibosiola D, Sastry N, Tyson G, Garimella K. Jettisoning Junk Messaging in the Era of End-to-End Encryption: A Case Study of WhatsApp. In Proceedings of the ACM Web Conference (WWW) 2022 Apr 25 (pp. 2582-2591). [ARI+22]

1.5.3 Other Publications in Collaboration

- Agarwal V, Vekaria Y, **Agarwal P**, Mahapatra S, Set S, Muthiah SB, Sastry N, Kourtellis N. Under the Spotlight: Web Tracking in Indian Partisan News Websites. In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2021 May 22 (Vol. 15, pp. 26-37). [AVA+21]

- Vekaria Y, Agarwal V, **Agarwal P**, Mahapatra S, Muthiah SB, Sastry N, Kourtellis N. Differential Tracking across Topical Webpages of Indian News Media. In Proceedings of the ACM Web Science Conference (WebSci) 2021 June 21 (pp. 299-308). [VAA+21]

- Beytia P. **Agarwal P**, Redi M, Kumar Singh V. Visual Gender Biases in Wikipedia: A Systematic Evaluation across the Ten Most Spoken Languages. Accepted for publication at the International AAAI Conference on Web and Social Media (ICWSM) 2022. [BARS22]

1.5.4 Dissertation Overview

The dissertation is organised as follows:

- Firstly in Chapter 2, I provide some more **Background** of political communication, interactions of various actors (*i.e.,* politicians, citizens and digital mass media platforms as shown in Figure 1.1) and challenges.

- Chapter 3 presents in depth study of the first type of engagement *i.e.,* **Interactions between politicians and citizens**. This includes attention dynamics, selective reply behaviours and tone of conversation. I present two case studies in this Chapter. The

first case study is on MPs active use of social media c.f. §3.7 and second on challenges such as online hate c.f. §3.8.

- Chapter 4 shows **Interactions between politicians with mass media**, including differential tracking by news based on users' personas (c.f. §4.7) and polarisation on Wikipedia (c.f. §4.8).
- Finally I present in Chapter 5 **Interactions between citizens with mass media**, indicating use of social media by multilingual users and supporters of political parties. With two case studies (c.f. §5.5 and c.f. §5.6), I characterise content mutation in multi-lingual user posts and misbehaviour in public chat groups respectively.
- In the end, I conclude with reflections on findings in this study, known limitations and a discussion on possible future of this research work.

CHAPTER 2

BACKGROUND

In this chapter I provide a wider background of the three forms of online conversations as discuss in previous chapter. Each section provides details of the broad theme and then discuss emerging challenges in that form of online conversation. Finally, I discuss how I build upon using these studies in later chapters.

2.1 Engagement between Politicians and Citizens

Although, political tweets might not look like substantive *contributions* to the political discourse, they appear to have become *an increasingly integrated element of political communication in a 'hybrid media system'* [Jun16]. In my research, I focus on *politicians and citizens* communications during periods when there is no election, and find that cross-party talk is more prevalent. Although, there was no election, I focused on a period full of political activity. For example, the EU Withdrawal and Implementation Bill (13 Nov 2017), a key piece of legislation on Brexit. This period also coincided with the rise of the #MeToo movement in Westminster, leading to a number of high profile resignations. I also chose a period that included days when Parliament was in session as well as in recess, making this a suitable period for a generalised study. However, the increase in use of social media platforms for political engagement has not only brought opportunities but also serious barriers to an open and deliberative public discourse. In following two sub-sections I provide more background on online interactions of politicians and citizens.

2.1.1 Use of Twitter in UK Politics

Earlier studies on the use of social media by UK politicians mostly relate to a period when such usage was in its infancy, with a small fraction of MPs being regular users of Twitter [JL11, LKM13, GBHVH13, GJB16]. Nevertheless, there were early indications that use of Twitter was entering the mainstream of electoral campaigning and political communications generally [BHG+20].

In earlier usage of Twitter, one-way communications ("broadcasting") predominated [GBHVH13, JL11, LKM13], but participatory communication through Twitter was seen to be emergent. It seemed to fit neatly into Coleman's [Col05a] concept of direct representation and politicians talked about their use of Twitter in these terms, but it was still secondary to other uses [Jun16]. Nevertheless, [GBHVH13] found that 19% of candidates' tweets during the 2010 general election campaign interacted in one way or another with voters, which they argued was a fairly substantial level of interaction compared to other forms of political communication during the campaign. A participatory style of communications on Twitter had potential to earn legislator's political capital [JL11] and was the only statistically significant strategy that had a positive impact on the size of the community [LKM13].

Cross-party communication *between MPs* was found to be unusual. Unsurprisingly, MPs indulged in one-off attacks on other politicians during the 2010 UK general election campaign [GBHVH13] but there was evidence of a more collaborative approach amongst an "organic community" of early adopters on Twitter [JL11]. The follower and followee, URLs sharing and other user actions (retweet and reply) on Twitter are used to study political ideologies of users [Bar15, LGG18, WTSC16]. Supporters of different parties tended to cluster around different hashtags, topics and politicians during election campaigns, creating politically separated communication spaces [Jun16, BJN+15].

2.1.2 Hate Speech, a Chilling Effect

Studies have shown that politicians face more online abuse and are subject to intense verbal attacks online [GGR+18, GBR+20, GVM+21, Sco19, FBB21]. Corroborating my results,

a study based on a manual annotation of 3000 tweets finds that male and female MPs both receive similar amounts of hate [WM20]. However, their qualitative methodology finds that the hate received by female MPs is more threatening. Gorell [GBR+20] finds that MPs who stood down at the 2019 UK General Election had more abuse than the ones who stood for election again.

A few studies have examined hate speech *by* politicians and the chilling effect this has had in other contexts, such as hate speech by politicians against Muslims [Pet19]. Rekker [RvS21] studied Geert Wilders' prosecution in the Netherlands and argued that his conviction eventually undermined democracy. [VSDV15] shows that Wilders' party's popularity increased as a result of the prosecution. This suggests that prosecuting or punishing politicians for hate speech can end up being counter-productive. Unfortunately hate speech by politicians can be extremely effective in changing public opinion, polarizing the electorate and increasing domestic terrorism [Pia20]. In spite of the challenges, research efforts are being made in the UK and other parts of the world to tackle the issue of online hate at scale [SB20, GGZY+20, dLPS+17]. Other efforts have helped further research by collecting and sharing datasets of hate speech in various contexts [DWMW17, FDC+18, dGPGPC18, QBL+19, Kag18, vARKL18, WH16, WTD17, ZMN+19, M+20]. In this research, I identify and study the "pile on" effect, whereby MPs during high workload times receive more hate as compared to other less busy periods for them. After identifying conversations with hate labels, I characterise the target of hate speech and the nature of cross-party in this conversation.

2.2 Engagement between Politicians and Digital Mass Media Platforms

It is of no doubt that internet, has seen a drastic shift in how citizens consume news and form opinions. More citizens cite online news sources as their primary source of news and other information than ever before[1]. In second part of my research, I study interactions between

[1] http://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/

politicians and digital mass media platforms with the use of content and profiling on news and online encyclopedia, *i.e.,* Wikipedia.

2.2.1 News Broadcasting and User Tracking

With most of the political discussion happening in the cyber-space, user experience leads the way of engagement and enchantment [MWWD06, ASW19]. At such a juncture, presence of news websites which promote unilateral partisan views of the facts has proven to be a real challenge to democracy and facilitated the rise of misinformation. At the same time, internet has facilitated the rise of another family of services, which tracks users using cookies and other browser fingerprints so as to profile them for advertising purposes. It is very valuable for the websites to profile their users, so as to allow personalised content.

Many studies have analysed the extent of online user tracking. In [KMBP18], authors measure the trackers per website and find that, on average, there are 10 different trackers per website; in news websites this average is 15 trackers. Similarly, in [EN16], authors used Alexa top 500 websites for each of the 16 basic content categories of Alexa. They find that websites providing editorial content embed the largest amount of trackers given their lack of external funding sources, and their need to monetise page views through advertising.

A popular web tracking mechanism is the use of cookies set on the user-side. In [HS19], researchers measured differences in third party cookies offloaded on a browser by collecting data from real users over a period of one year, before and after the GDPR regulation [Eur18], and found that the number of cookies does not change, not least because users tend to pay little attention to the GDPR options. This does not exclude sites that deal with news or political content. But one would expect that the amount of tracking a user on these sites is exposed to would be uniform. Another selective political initiative is the highly polarized HPWs in US politics, studied by [BJBS18]. In this research I study how after HPWs declaring themselves right- or left-leaning in their web portals such websites still exist and track different users.

2.2.2 User Personalisation and Polarisation

News personalisation is a way to send curated news briefing to subscribers [MW14, TS12]. Studies have tried to understand the political stance of users and news websites by various methods. Examples include the use of NMF clustering ($k = 2$) [LGG18], finding non-binary political ideological scores, sentiment analysis [PH14], twitter profile data [PP11], etc. In [ERE+15], authors simulate users browsing the web in an attempt to study the possibility of a passive eavesdropper leveraging third party HTTP tracking cookies for mass monitoring. They find that indeed the eavesdropper can reconstruct 62-73% of a typical user's browsing history. In [KHN+18], authors studied 5 million paid ads on Facebook and claim that US users are mostly targeted by foreign groups using loopholes in regulations of Facebook.

Taking a recent case of Politicians engagement on Wikipedia, a platform where thousands of volunteers revise and add content constantly [INPG10], this research covers first study looking at political communication and polarisation in the context of UK Members of Parliament. There are several works on studying Wikipedia content across various dimensions. Certain editors proclaim their political leaning and form communities [NLK+13]. Building on previous works I quantify contributions by polarised teams. Those consisting of a balanced set of politically diverse editors may create articles of higher quality than politically homogeneous teams [STDE19].

2.3 Engagement between Citizens and Digital Mass Media Platforms

Finally, in this part I study *citizens using digital mass media platforms* for political discourse. My research focuses on multi-lingual political discourse in the largest democracy in the world, India with usage of platforms like *ShareChat* and *WhatsApp*. In the following two sub-sections I present background of these platforms and previous studies in this field.

2.3.1 Multilingual Users and Social Media

There have been a number of works looking into the multi-lingual nature of certain platforms. For example, language differences amongst Wikipedia editors [KS18], as well as general differences amongst language editions [Hal14, HG10, Hal15]. Often this goes beyond textual differences, showing that even images differ amongst language editions [HLAH18]. There have also been efforts to bridge these gaps [BHC+12], although often differences extend beyond language to include differences between facts and foci. This often leads editors to work on their non-native language editions [Hal15]. Although highly relevant, my work differs as I focus on a social community platform in which individuals and groups communicate. This contrasts starkly with the use of community-driven online encyclopedias, which are focused on communal curation of knowledge.

More closely related are the range of studies looking at the multi-lingual use of social platforms, *e.g.,* blogs [Hal12], reviews [Hal16] and Twitter [NTC15, HCC11]. In most cases, these studies find a dominance of English language material. For example, Hong *et al.* [HCC11] found that 51% of tweets are in English. Curiously, it has been found that individuals often change their language choices to reflect different positions or opinions [MBGM15]. Another major part of this interaction is the study of memes and image sharing. Recent works have looked at meme sharing on platforms like Reddit, 4Chan and Gab, where they are often used to spread hate and fake news [ZCB+18]. These modalities are powerful, in that they often transcend languages, and have sometimes provoked real world consequences [ZCS+19] or have been used as a mode of information warfare [ZCDC+19, RJC+19].

2.3.2 Misuse by Sending Unwanted and Unsolicited Messages

Although there have been extensive studies of spam on social media [SKV10, G+10, YRS+10], email [Cor08] and SMS [P+19], spam has not yet been studied on WhatsApp at scale, except for anecdotal observations [TR20, AS19, Pat19b].

There have been prior studies on the misuse of messaging services such as SMS fraud in Pakistan [P+19]. There are also a small set of related works looking specifically at

the dissemination of malicious content via WhatsApp groups. These, however, are so far focused on misinformation. Resende et al. [RMS+19] investigated the dissemination of misinformation during the Brazilian elections. The authors also studied the impact of introducing limits on message forwarding [dFM+19]. In another similar study of Brazilian elections, Victor *et al.* [BB19] find partisan activities in these political groups. In the case of WhatsApp , this can be problematic due to its use of end-to-end encryption. Hence, Reis [RMGB20] proposed an architecture to flag misinformation in WhatsApp without breaking end-to-end encryption. Unlike approach present in this research, [RMGB20] relies on a manually annotated set of image hashes.

2.4 Discussion

While many computational research efforts rely solely on processes of quantification, policy researchers and the legal discipline have been taking a more qualitative approach. My research differs substantially from the works above and helps in quantifying new form of digital citizen engagements among Politicians, Citizens and Digital Mass Media Platforms. With a focus on a period without an election also makes my efforts complementary to the large number of works that examine the Twitter during election campaigns. Also, a focus on emerging online user language communities which are formed naturally helps in deeper understating of language with strict distinctions between communities. Furthermore, with new usage of technology comes some underlying challenges. I study some of those challenges as a part of this research.

CHAPTER 3

DIGITAL ENGAGEMENT INVOLVING POLITICIANS AND CITIZENS

In a democracy the poor will have more power than the rich, because there are more of them, the will of the majority is supreme.

Aristotle

3.1 Introduction to Political Engagement

In this chapter, I characterise how engagement of democratic representatives with their citizens is shaped by online platforms, more specifically Twitter. I focus on the UK, where a remarkable 579 out of 650 Members of Parliament (MPs) are active on Twitter. To frame the discussion, I consider the nearest offline equivalent for interaction between MPs and citizens – *constituency service*. The traditional means by which this is done is for the elected representatives to hold open and private meetings with those that elected them. In the UK, for example, MPs travel back to their constituencies, typically on Thursdays, after the work of the Parliament is done, and hold 'surgeries' with their constituents. 'Town hall' meetings in the USA serve a similar purpose. Constituents may also phone or email their MPs and members of Congress to let them know their positions on key issues.

To be sure, there are differences between engagement on Twitter and traditional constituency service. Twitter interaction can be immediate and spontaneous, in contrast with scheduled surgeries. The public nature of Twitter renders it unsuitable for constituency service requiring

personal information. In my collected dataset in 1548 cases, MPs asked to move away from public discussions on Twitter, asking constituents to make appointments at their surgeries, offering their email addresses or asking the respondent to "DM" or "direct message" them. In a handful (≈ 10) of cases, both options were offered[1]. These interactions represent over 7% of replies by MPs. Also, constituency service is usually seen as MPs engaging with and serving those in the *geographic area they represent*. In the UK, there is even a strict parliamentary protocol that MPs do not seek to intervene or act in matters raised by the constituents of other members [IRI10]. On Twitter, however, it can be hard for MPs to tell the precise location of their correspondents, and the immediacy and public nature of the medium may lead to interactions with non-constituents.

Despite such differences, both forms of communications hold the same promise: direct contact and engagement between elected officials and those they are supposed to represent. Therefore, I turn to the literature on constituency service interactions for pointers on the nature of the discourse between citizens and UK MPs on Twitter. The traditional view of psephologists has been that constituency service is worth only about 500 votes [NW90, BC01], and thus, is insufficient to make a difference in all but the closest of elections. [Kin91, Kra97] talk about the 'incumbency factor' and the need for MPs to develop this relationship in order to get re-elected. Thus, more than being a campaigning tool, engaging with citizens can be seen as a mechanism for building relationships and achieving better representation.

Based on these considerations, I focus on a two-month period from Oct 1, 2017 to Nov 29, 2017, when there was no election going on, and thereby seek to understand the usage of Twitter as a tool for everyday citizen engagement. The period also encompasses times when Parliament was in session (requiring MPs to be away from their constituencies, attending the House of Commons) and in recess[2] (when MPs are free to return to their constituencies), and therefore can be expected to cover both aspects of MP activities. I study the Tweets, Retweets and Replies of MPs towards other users, as well as from other users towards MPs.

[1] One MP wrote: "@XX, If you follow me I'll DM you or please email YY@ZZ and I will get back to you. Thanks, D"

[2] https://www.parliament.uk/about/faqs/house-of-commons-faqs/business-faq-page/recess-dates/

Both groups are active, with the MPs and the citizens respectively producing 178,121 and 2,339,898 Tweets, Retweets and Replies directed at each other.

The parallels and distinctions between engagement on Twitter and constituency work also drive my research questions: Norton and Wood's seminal study of British MPs' constituency work in the 80s concluded that constituency service can be extremely rewarding, although taxing, taking MPs close to "saturation point" [NW93]. Therefore it is natural to ask whether Twitter imposes a burden on MPs. Given that any additional work would also likely take time away from other duties of the MP, I also wish to understand how MPs manage whatever burden is imposed on them by their Twitter presence. Following the typology of [Sta08] who studied how MPs use constituency service to package themselves, I ask whether MPs are using Twitter to prioritise helping constituency members, gaining personal visibility by highlighting work they have done, for spreading the message of their party and party leaders, or for other purposes. Finally, I am interested in identifying the tone of the conversation online. Given the tendency of Twitter as a polarising and sometimes aggressive sphere [CKB+17, CRF+11], I ask what the nature and tone of the conversation is, between MPs and others. This discussion can be crystallised into the following research questions:

1. As a new and additional medium of citizen engagement, how much load does Twitter place on MPs, and how does this load vary?

2. Do MPs selectively prioritise certain forms of engagement or seek external help (e.g., from their staff)?

3. What is the nature and tone of the conversations? Is Twitter a polarising sphere with echo chambers for each party and side of the political spectrum? Is the tone civil or aggressive?

I find that attention to individual MPs varies dynamically: although there is a huge amount of information overload during short periods of time which I term as "focus windows", there is a significant amount of "churn" in the set of MPs who are "in focus" at any given time. MPs strategically manage their relationship with their followers and this information overload by

balancing their different roles as representatives of their constituency and their party [Sta08]: They selectively reply to Twitter profiles in the UK and within their constituency region, fulfilling their representative role, and use retweets as a mechanism to promote their image and spread the message of their party. Interestingly, I find evidence of significant cross-party interaction, between citizens who support and follow MPs from one party, and MPs from other parties. Thus, in an atmosphere of growing political divide in the UK (e.g., [JCG17, SB17]), Twitter seems to offer ways to avoid the "echo chamber" behaviour which characterises much consumption of information about politics online.

3.2 Background of the UK Parliament: MPs and Democracy in the UK

The Parliament in Westminster is the supreme legislature in the UK. It is composed of two houses or chambers. The primary house is the House of Commons. It has 650 elected members. Most MPs at any given election are drawn from a handful of major political parties. It is possible for candidates to run for election without the backing of a political party, but they are very unlikely to get elected. The three major parties in the UK Parliament following the 2017 general election were Conservative (Cons.), Labour (Lab.) and Scottish National Party (SNP). Other parties with MPs include the Liberal Democrats (Lib Dems) and the Democratic Unionist Party (DUP).

Members of Parliament are elected on a "party ticket" or manifesto and when they vote in the House of Commons, they are expected to obey party discipline. This also applies to their publicity and engagement work, where they are discouraged from giving messages that are inconsistent with the party line. This role of the MP as a party representative may sometimes conflict with the role of MPs as representatives of their constituencies. However, some MPs are more loyal to their party than others [Cow02], and in some cases, may choose constituency over party.

3.3 Datasets

In this section, I focus MPs and their communication needs and motivations, and describe the datasets which I have collected to answer research questions[3].

The dataset used for the analysis in this chapter consists of 2.5 Million tweets covering the period from 1 October 2017 to 29 November 2017. A number of considerations led to this particular period as a choice. Since, I also wanted to understand online hate, I focused on a period during which the EU Withdrawal and Implementation Bill was introduced (13 Nov 2017), as this was a key piece of legislation on Brexit during a highly fractious period in British politics. This is an unusual time in the UK Politics where a lot of attention is there to politicians and the policies they are introducing in the UK Parliament. This period also coincided with the blow up of the #MeToo movement in Westminster, leading to a number of high-profile resignations. I chose a period that included days when Parliament was in session as well as in recess. There were other events and scandals that happened within this period (e.g. the resignation of Priti Patel as International Development Secretary). Note that other periods of study can indicate a different level of engagement and hence hate towards MPs. In my research, I collected a dataset to cover various aspect of online political engagement in the UK. I wanted a time period that was far enough in the past so that it did not cloud annotator judgement yet is close enough that there was sufficient online activity by MPs and citizens, including evidence of online hate. So as not to influence or be influenced by current politics, I wished to choose a time period before the current premiership of Boris Johnson.

During the period chosen, no party held an overall majority of seats in the House of Commons. The Conservative Party was the largest party with 317 seats, followed by the Labour Party (262), the Scottish National Party (35), the Liberal Democrats (12), the Democratic Unionist Party (10), Sinn Féin (7), Plaid Cymru (4), and the Green Party (1). The remaining two seats were held by Lady Sylvia Hermon, an independent MP in Northern Ireland, and the House of Commons Speaker, John Bercow (by convention the Speaker

[3]The dataset is made available at https://bit.ly/2HSTOsa for research usage. Following Twitter's Terms of Service (https://twitter.com/en/tos), only Tweet IDs are shared.

severs all party ties during their time in office). Following a general election held on 8 June 2017, the Conservative Party formed a minority government, relying on the support of the Democratic Unionist Party to achieve a governing majority through a 'confidence and supply' agreement.

MPs on Twitter I start with the 579 MPs who are active on Twitter, as obtained from MPsOnTwitter website[4]. The UK House of Commons has 650 Members of Parliament (MPs). Of these, 579[5] MPs (187, i.e., 32.37% are female) are active on Twitter, with a total of $\approx$ 13 Million followers. For each MP, I obtained the following data:

Follower and Following Using Twitter API, I fetched all the users ($\approx$ 4.28 Million) who follow MPs and also the users that MPs followed (869K).

Demographics of MPs Data on the characteristics of MPs can be downloaded from the Members Names' Information Service (MNIS) [Par13] API. This is one of several public APIs maintained by the UK Parliament that records the work of Members. MNIS contains data on each Member's name, gender, constituency, political party, and on any government or opposition roles they have held — these are ministerial positions in the government, or equivalent roles in opposition parties' front bench teams[6]. One characteristic of interest that is not held in MNIS is an MP's ethnic group. The think tank British Future compiles some data on the ethnicity of MPs in order to assess the extent to which the ethnic composition of the House of Commons reflects the society it seeks to represent. British Future has said that in compiling these lists they "follow a liberal principle of self-definition, so that where candidates define themselves as being from ethnic minority or mixed heritage backgrounds in their own public statements, they have been included in these figures" [KB15, Fut17].

[4] http://www.mpsontwitter.co.uk/list

[5] I have collected data for 559 MPs since last year, and have not included the 20 new MPs who have joined since then.

[6] Four MPs left their political parties and became independent during the period of analysis. The dataset shows their party at the start of the period, before they became independent.

MPs on Twitter	579
Verified MPs	83.54%
Total Followers	12.83 Million
Mentions per MP per day (⇐)	Mean: 67.96, Median: 13.25, SD: 302.82
Activity (Tweets, replies to others) per MP per day (⇒)	Mean: 5.52, Median: 3.3, SD: 6.18

Table 3.1 Details of the Twitter dataset (Oct 1 – Nov 30 2017).

Tweets and replies (⇒) Using the Twitter API, I obtain from the MPs' Twitter timelines a total of up to 3,200 original tweets, retweets and replies to other Twitter handles. This covers the period of Oct 1 - Nov 30 2017, and I am able to fetch all MPs' timelines within the maximum limit of 3,200 allowed by Twitter. I verify that for all MPs the date of first post in 3,200 posts is prior to start date of my study period. I discard those posts by MPs which are outside Oct 1 - Nov 30 2017 period. These collectively identify the utterances made *by* the MPs, directed *towards other* Twitter users. I will use the symbol ⇒ to refer to such Tweets.

Mentions and replies to MPs (⇐) To fully understand the extent of the conversation, I obtain the utterances of all *other* Twitter users[7], directed *towards* the MPs. This is obtained by searching for the MPs' Twitter handles using Twitter's "advanced search" API, and includes all mentions of the MPs' Twitter handles, whether as a reply to a tweet of an MP, or merely mentioning an MP's Twitter username in a non-reply Tweet. I will use the symbol ⇐ to refer to such Tweets.

Collectively, ⇒ and ⇐ capture both sides of the conversation between MPs with Twitter handles and the rest of Twitter. To understand how people talk about MPs *who are not on Twitter*, I searched for the full (first and last) names of such MPs, obtaining 35,904

[7]In the rest of this chapter, I interchangeably use the terms "ordinary" Twitter users and "citizen" to refer users who have mentioned an MP in one of their Tweets. When the term citizen is used, it has been verified (if relevant), that the users included are those who declare a profile location in the UK (See Geography details above).

Tweets. To ensure that this refers to MPs and not some other person with the same name, I manually examined all Tweets and compiled a list of the most politically related words used in conjunction with these names (`mp, Brexit, Parliament, Westminster, Tory, Minister, Party, Conservative, Labour, Vote, Democracy`). Filtering for these keywords, I am able to retain 15,083 (of the 35,904) Tweets. Since these are Tweets mentioning MP names, I term these as "*pseudo-mentions*". Clearly, this is being conservative in capturing pseudo mentions, and may have ignored several tweets, e.g., those that may not use the first and last names of the MPs. However, pseudo-mentions introduce data about 67 MPs. Thus, information about *646 of the 650 MPs are captured in my dataset, in one way or another.* Statistics about the dataset are listed in Table 3.1.

I also perform additional heuristic processing to obtain the following information for MPs and all users who mention them or are mentioned by the MPs (through retweets, replies or original tweets). Where heuristics may lead to errors, I try to perform some checks, using limited ground truth to give some indication of the accuracy or coverage that I believe I have attained:

Geography I assign country-level labels for each user as follows: Fetching the profile location of users, and checking for words such as UK, London, England, Wales, Scotland, Northern Ireland, United Kingdom, I labelled around half of the users. For the remaining ones, the unique list of their locations is passed to Photon library's geocode function[8], an open street API to label countries given location names (e.g., city/locality/state etc.) that a user might have used. Using this procedure, the countries of $\approx$ 82.7% (77.3%) of users retweeted (replied to) by MPs were obtained. Note that where I see ambiguity in place names I choose the location within the UK. This assumption is based on the fact that users who follow MPs tend to be from their own country.

For users in the UK, I dig deeper. A Twitter profile location may mention only a city name (such as 'Cambridge, UK'), rather than a specific constituency (such as 'Cambridge South', which contains a small part of the City of Cambridge and close-by

[8]http://photon.komoot.de/

villages). Also, citizens may work in one constituency and reside in another which is close by. Thus, to identify interactions between a user and their local MP, I translate user locations to the 'postcode area' which represents the region (e.g., 'CB' represents Cambridge in Cambridge-related postcodes), and consider all interactions between users from that postcode area, and the MPs representing that postcode as interactions between MP and a potential constituent. I fetch these postcodes area again using Photon library's geocode function based on location name. The Photon API returns the geometry coordinates of the center of the location[9]. I use this center point coordinates to obtain a postcode. From the postcode I extract those characters which represent postcode area of that location.

Party affiliation For every user who has mentioned an MP, I associate the party affiliation of the MP with the user. Users have followed a mean (median) of 3 (1) MPs, from an average (median) of 2 (1) parties. I affiliate each user with one party. For users who have followed MPs from more than one party, I assign them the party they have followed the most following an early study [Bar15]. Note that a user can follow MPs due to multiple reasons [KE19]. A deeper study using the dataset in this research can be useful to understand why users follow and unfollow MPs. In this research, the two main parties (Conservative and Labour) have nearly the same number of MPs. *Check:* By checking for the presence of party names in the profile description amongst a sample of $\approx$ 8K Conservative and Labour supporters, I am able to correctly label nearly 7,400, yielding at least 91.4% accuracy for my heuristic.

Hate speech annotations There have been a number of hate speech-related models developed in recent years [DWMW17, WTD17, GGR$^+$18, M$^+$20]. Each of them has slightly different definitions of hate and may be trained on data from different contexts and platforms, which in turn has measurable effects on what is labelled as hate [M$^+$20]. To examine the effects of different training sets and models on my datasets, I use 18 different hate speech classifiers. These are ultimately based on two widely used

[9]Example response for city 'Cambridge': https://photon.komoot.io/api/?q=cambridge&&limit=1

models developed by Wulczyn *et al.* [WTD17] and Davidson *et al.* [DWMW17]. Each model is trained on on 9 different publicly available datasets:Davidson [DWMW17], Founta [FDC+18], Gilbert [dGPGPC18], Jing-Gab [QBL+19], Jing-Reddit [QBL+19], Kaggle [Kag18], Wazeem [WH16], Wulcyzn [WTD17] and Zampieri [ZMN+19] (Check Appendix Section 7.1 for more details of each dataset). This yields 9 x 2 = 18 variants. The intuition is that if a large number of these 18 classifiers consider a tweet as hateful, it is likely to be "truly" hateful. Thus I take the majority across the 18 classifiers. Figure 3.1 measures how the majority vote of the 18 classifiers performs, in two ways. On the x-axis, tweets are binned by the number of classifiers n that label those tweets as hateful. As expected, with increasing n, the volume of tweets labelled as hateful by n of the 18 classifiers decreases. I also measure the average 'toxicity' of the tweets in each bin, according to toxicity scores collected using parallel calls to Google Perspective API [API21]. Note that the exact definition of toxicity could differ from the definition of hate speech. I compute average toxicity of tweets in different bins as a proxy to understand hate speech as toxic comments towards MPs. Again, as expected, the average toxicity is higher in bins which contain tweets that a larger number n of classifiers agree as hateful. Interestingly, for $n > 9$, i.e., in bins which contain tweets that a majority of the 18 total classifiers are agreed that the tweets are hateful, I find that the toxicity level is higher than 0.8, the recommended toxicity score to consider a tweet as hateful [Med21]. Furthermore, approximately 1% of the total volume of tweets can be found in bins $n > 9$, which also corresponds to known volumes of hate speech on Twitter [PKQSLCC19]. Based on these considerations, in my study, I consider tweets in bins $n > 9$ as the 'hateful' tweets. Note this choice of n will lead to a high precision, but low recall hate speech dataset. This means that the volume of tweets labelled as hateful will certainly fall under the hate speech category, though it might miss-classify hateful and label them as non-hateful. Future studies can build more accurate models using the open dataset from this study.

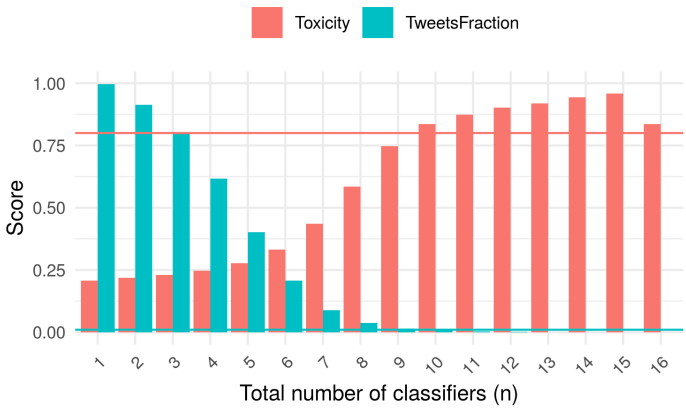


Fig. 3.1 Classifier Agreement vs. volumes of tweets they agree on and average 'toxicity' of those tweets. Each tweet is binned into one of 18 bins based on the number n of classifiers that label the tweet as 'hateful'. As n increases, more classifiers agree that the tweets in the bin are hateful. Only 16 bins are used as no tweet is classified as hateful by more than 16 classifiers. The volume of the tweets in each bin (coloured as cyan) decreases with n. However, the average 'toxicity' score in each bin (coloured as red) increases with n. A red horizontal line indicates the default toxicity score of 0.8 recommended by Google Perspective API for labelling hateful content and the horizontal cyan line (close to x axis) represents a volume of 1% of tweets.

3.4 Dynamics of Citizen Attention

In this section, I approach the first research question, and estimate the burden caused to MPs by their Twitter presence, by studying tweets directed *towards* MPs by other Twitter users. My starting point is the stark difference in Table 3.1 between the number of mentions that an MP gets (marked as $\Leftarrow$), and the average number of tweets and reply activities made by them (marked as $\Rightarrow$). This suggests that MPs could be overloaded, and are not able to respond to all tweets directed at them.

To examine this, I introduce metrics that measure the spread of attention load, in terms of mentions of MPs. I study the distribution of attention across time for any individual MP, and across all MPs during any given time window. I find that in any given time window, a small number of MPs are 'in focus', and receive a large number of mentions. However, as the news cycle moves on, other MPs' activities come into focus. I then illustrate this phenomenon using examples and discuss the implications.

3.4.1 Attention is Focused during Small Time Windows

To understand how overloaded the MPs are with the number of mentions they receive, I first examine high activity periods. I define a period of *high activity* as a continuous sequence of days when the daily activity is considered as 'high'. Formally, given an MP i and a threshold average number of mentions T_i, I define a continuous sequence of days R as a *high activity window* for MP i if it satisfies the property

$$high\_activity_i(R) : \sum_{d \in R} v_{id} > T_i|R|$$

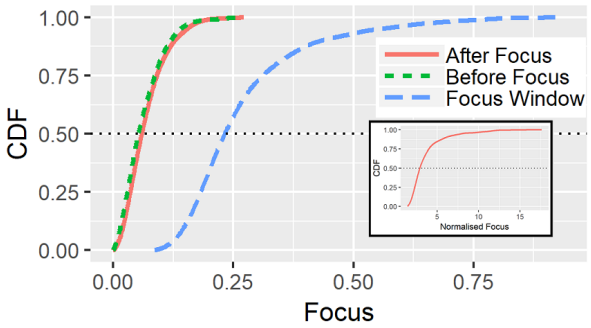
where v_{id} is the number of mentions obtained by MP i on day d.

In this study, I set the threshold for a high activity individually for each MP. A day qualifies as a 'high activity' day for an MP if the number of mentions received by the MP that day is higher than the *personal* average for that MP[10]. Note that even if MPs only receive a large number of tweets during a short time window, this will increase their personal average for the whole 2 month duration of my dataset. I term the longest continuous run of days during which an MP i has more than his or her personal average number of Tweets mentioning them – as their *focus window* $T_i^{\max}$. I can compute the fraction $F(T)$ of MP i's mentions that are obtained during a time window T as

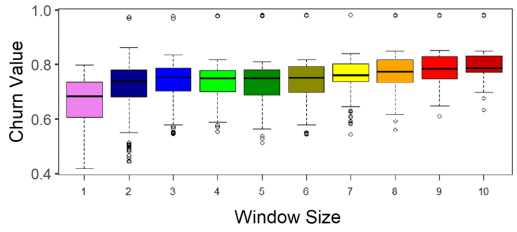
$$F(T) = \frac{1}{V_i} \sum_{d \in T} v_{id}.$$

Here, V_i is the total volume of Tweets mentioning MP i in the dataset, and v_{id} is the number of mentions obtained on day d in the window T. I define the *Focus* of MP i as the fraction of mentions $F_i = F(T_i^{max})$ obtained during the focus window $T_i^{\max}$. In other words, Focus measures what fraction of an MP's mentions during the whole 2 months period covered by my dataset is concentrated during the small Focus Window, *i.e.,* the longest continuous sequence of days during which the MP receives a higher than average number of Tweets.

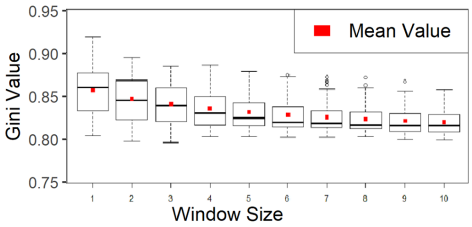
[10]Other threshold definitions were examined, but not reported here due to space.



(a) Focus



(b) Churn



(c) Gini

Fig. 3.2 The distribution of (a) **Focus** of MPs' mentions: CDF (Cumulative Distribution Function) of fraction of mentions $F(T_i^{max})$ obtained by MPs during their focus windows (Blue), just before focus window (Green) and just after after the focus windows (Red). This clearly shows that before and after focus window the fraction of mentions are much smaller. Inset: CDF of normalised Focus. Most of the mass is several times over 1, confirming high information overload during focus windows than average. (b) **Churn**: Box plots of the distribution of Churn values across time windows of different sizes. The box extends from the lower to upper quartile values of the data, with a line at the median. The mean is shown as a green dot. Whiskers extend from the box to show the range. Flier points are outliers past the end of the whiskers. (c) **Gini**: Box plot of distribution of Gini co-efficients of the number of mentions received by MPs during time windows of different sizes.

Figure 3.2a shows the CDF (Cumulative Distribution Function) of Focus values for all MPs. For comparison, the fraction of mentions $F(T_i^b)$ and $F(T_i^a)$ during similar-sized windows before and after the focus window is also shown. Mentions tend to fall off rapidly outside focus windows: in the windows immediately preceding (following) these periods of intense activity, MPs on average receive less than a quarter of the tweets received during the high activity focus window.

By definition, Focus takes values between 0 and 1. To get a sense of how skewed focus are values, I normalise the obtained focus based on the expected fraction of mentions given the size of the focus window: if the V_i mentions are evenly across a total of D days, the

number of mentions expected in a focus window of $|R|$ days is simply $|R|/D$. The observed Focus F_i can therefore be normalised as $F_i D/|R|$. If mentions are uniformly distributed, normalised focus would be ≈ 1. Figure 3.2a (Inset) shows that this value tends to be several times larger than 1, suggesting that a disproportionately large fraction of the mentions for an MP might come during their one concentrated focus period. In other words, MPs are in the limelight only for a short period of time. Empirically, I find that the focus window period typically lasts between 3–5 days.

3.4.2 Attention is Unequal but Focus Moves among MPs

The previous discussion suggests that MPs receive mentions in a very bursty manner: Outside their focus window, an individual MP contributes much less to the overall volume of mentions directed towards MPs. Yet, as seen earlier, there is an average daily volume amounting to about 68 mentions per MP. In this subsection, I look at how mentions are shared among the MPs on a daily basis.

I proceed by considering all possible time windows of different sizes from 1–10 days. For instance, I can have 5-day windows from Oct 1–5, Oct 2–6 ... Nov 24–29. The goal is to understand the effect of MPs not receiving many mentions outside their focus windows and how mentions are shared during any given window.

For any time window of a given size, I ask how many of the high activity MPs of that window – MPs who receive more than their personal average number of mentions – continue to receive high numbers of mentions in the next window. Formally, I define the set of active MPs during a time window R as

$$active(R) = \{i | high\_activity_i(R)\}.$$

I can define the *churn* of a time window R and the time window R^+ immediately following it as the difference in the set of active users between the two windows:

$$Churn(R) = \frac{|active(R) \triangle active(R+)|}{|active(R) \cup active(R+)|}$$

where the numerator is the symmetric set difference of MPs who are active in time window R but not R^+ and vice versa, and the denominator is the union of users in the time windows. Figure 3.2b shows that churn is high: Nearly 70% of MPs who receive more than their personal average of mentions during one window are not able to sustain this level of activity in the next window. Churn increases slightly as window sizes increase, with MPs finding it difficult to continuously receive high numbers of mentions over larger time windows.

While churn looks at differences across time windows, I also measure how unequal the attention distribution is *within* a time window, by counting the number of mentions each MP receives during the window and computing the gini co-efficient [Cow00] across all MPs receiving mentions[11]. Taking an MP i, the following equation computes Gini coefficient for a specific time window based on mentions towards each MP x_i and n as the total number of MPs.

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \overline{x}}$$

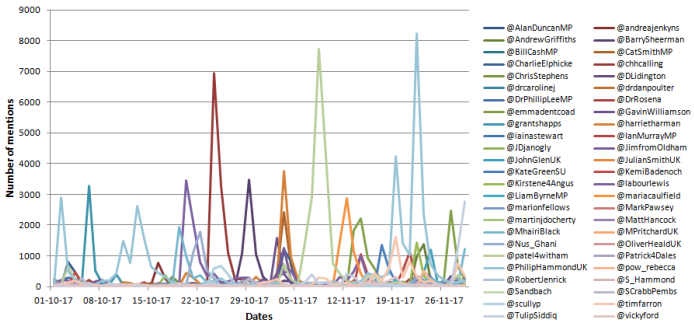
Gini co-efficients vary between 0.8 and 0.92, and the closer it is to 1, the more unequal the distribution being measured is. I can get a sense of the inequality for different time windows of a given size by looking at the distribution of Gini co-efficients. Figure 3.2c shows the distributions of gini co-efficients for time windows of different sizes. The median gini co-efficient for all window sizes is consistently above 0.8, indicating that in any single window, most of the mentions are for a small minority of "attention rich" MPs.

Collectively, these results suggest that during any given window, a few MPs are attention rich and receive a large number of mentions, but this set of MPs shifts over a period of days, so there are no overall "superstars" who are always at the centre of attention. The examples below serve to illustrate this phenomenon.

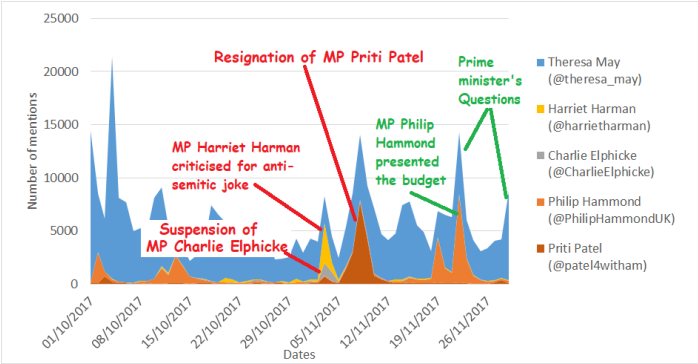
3.4.3 Examples and Implications

To better visualise the attention imbalance, Fig. 3.3a plots the daily mentions volumes of 50 MPs with the highest focus. The focus values for each of these MPs is more than 0.5; i.e.,

[11] Using R package Gini https://www.rdocumentation.org/packages/DescTools/versions/0.99.45/topics/Gini



(a) MPs with highest focus



(b) Examples of attention

Fig. 3.3 Timeline of number of mentions (a) 50 MPs who have more than half their mentions occur during their focus window. (b) Examples of (un)anticipated attention towards MPs in green (red) with label of the corresponding events.

more than half of their mentions were received during their focus windows. The spiky nature of the graph illustrates how attention can be highly concentrated during short focus windows (typically 3–5 days), and moves on to other MPs after the focus period.

Priti Patel (@patel4witham) represents an interesting example (Fig. 3.3b): She was International Development Minister until 8 Nov 2017, but was forced to resign as a result of a scandal caused by unofficial meetings with Israeli ministers while on a holiday in that country. This resulted in a barrage of focused attention which fizzled out as other new stories cropped up. Similarly, Philip Hammond (@PhilipHammondUK), Chancellor of the Exchequer, had a huge number of mentions around the Autumn Budget (22 Nov 2017, Fig. 3.3b). Note that there is a smaller spike for Hammond just before the budget when he mistakenly claimed in an interview that there are no unemployed in the UK[12].

[12]theguardian.com/politics/2017/nov/19/there-are-no-unemployed-in-uk-says-philip-hammond-tv-gaffe

These two examples illustrate two different kinds of focus windows: The attention towards Priti Patel was completely *unanticipated* until the event unfolded, whereas the increased attention towards Philip Hammond as he presented the budget was predictable and could have been *anticipated* and planned for (although even here, unanticipated mistakes can create spikes, as in Hammond's case).

Anticipated attention is mostly for positive events and is in many cases "manufactured" by the MPs, their staff and members of their party, following prominent speeches or comments made in Parliament, as such successes are advertised by sharing widely on Twitter. A common source for such high attention events is activity during Prime Minster's Questions, which happens every single Wednesday at noon when the House of Commons is in session, and usually involves a lively and sometimes raucous debate. When an MP makes a particularly valuable (or sometimes particularly witty) contribution, it is shared by the MPs themselves, or by others, on Twitter, and then gets widely discussed.

By contrast, unanticipated focus windows, as with Priti Patel, include mostly negative events for the MP (see Fig. 3.3b). For instance, during the Westminster sex scandal, Charlie Elphicke (@CharlieElphicke), a Conservative MP, was accused of sexual misconduct and subsequently suspended from his party. The Westminster sex scandal, which coincided with the '#MeToo' movement, includes several other resignations and castigations which also received high attention and focus values. Similarly, Labour MP Harriet Harman (@HarrietHarman) was criticised for mentioning an anti-semitic joke on live TV.

The focus windows of 70% (35/50) of the MPs in Figure 3.3a are for events that could have been anticipated. However, perhaps unsurprisingly, unanticipated windows receive unusually high attention – four of the top *five* focus windows are apparently unanticipated, and for events which generated considerable adverse publicity. Thus, unanticipated attention can be all the more difficult to manage because of the volumes. Furthermore, all of the top five focus values are for MPs from the Conservative Party, which, as the current ruling party, tends to receive a large amount of scrutiny. Four of these also had ministerial level roles at one point or another and another held a senior role within the party. The fact that even such

prominent MPs obtain more than half of their mentions during a small 3–5 day focus period illustrates that the attention of citizens is highly volatile and all too brief.

Focus periods represent opportunities for the MPs to raise their profile and engage with the populace on issues important to the MP. Whether the focus is a result of a positive event that the MP can take advantage of, or a negative event the MP should defend against, being able to appropriately handle the situation and manage the (brief) attention overload is critical. The next section looks at strategies that MPs use to manage citizens' attention both during their focus periods and out of their focus periods.

3.5 Managing Citizens' Attention

Incoming Tweets mentioning MPs (marked as $\Leftarrow$ in Table 3.1) can be seen as a means for UK citizens and other Twitterati to engage with the MPs. In the previous section, I established that MPs faced an information overload with incoming tweets, especially during focus windows. In this section, I turn to the second research question, and ask how MPs manage this attention load in responding back, i.e., I also take into account the MPs' outgoing Tweets (marked $\Rightarrow$) in terms of Tweets, Retweets and Replies, and ask how MPs engage with the rest of Twitter.

I identify two possible adaptations: The first consists of very selective replies, with MPs prioritising interactions with users local to their constituency region. The second is to employ staff who can help manage the load. I find extensive usage of the first strategy, with MPs largely prioritising their responses to users local to their region and to UK users. However, only some MPs appear to be using additional staff who can help manage their social media profiles.

3.5.1 Selective Replies and Localism in MP Actions

MPs tend to be very busy, and being active online takes time away from their other duties, and their real-world constituency [Jac08]. Therefore, the expectation is that MPs would be selective in who they respond to (even during non-focus periods). I test this hypothesis in

Geography	%Mentions ⇐	%Retweet ⇒	%Reply ⇒
UK (Constituency (C))	74.16 (C:28.8)	90.39 (C:56.7)	89.93 (C:59.04)
Commonwealth	3.88	2.35	1.74
USA	5.34	3.42	4.63
EU	3.47	2.53	1.72
Others	13.15	1.29	2.06
Total	100%	100%	100%

Table 3.2 Geographic Distribution of incoming mentions (⇐) and outgoing actions (⇒) in different geographical regions. Each column adds up to a whole (i.e., UK + Commonwealth (British Commonwealth and Overseas Territories) + USA + EU + Others = 100%). UK replies are further subdivided into replies within the constituency region of the MP, and those outside (The percentage local to MPs' constituency regions is shown in parenthesis as C:XX.YY%). Thus, for instance, from (Row 1, Col 1), 74.16% of all incoming (⇐) mentions towards an MP come from within the UK. **Of these**, 28.8% of mentions are from within each MP's constituency, and the remaining (71.2%) are from outside their constituencies.

two ways. First, I check the geographic areas of those whom the MPs respond to. Next, I check the category of the Twitter handles they respond to – whether they are responding to other MPs, or those that they follow or are following them.

Table 3.2 shows the percentage distribution of the incoming mentions (⇐) and outgoing actions (⇒) – replies and retweets, among different geographic regions. In their conversations with Twitter users not from the UK, MPs tend to favour responding to Twitter users from countries that the UK has ties with: USA (5.3% mentions, 3.4% retweets, 4.6% replies), British commonwealth and Overseas Territories (3.8% mentions, 2.3% retweets, 1.7% replies) and the EU (3.4% mentions, 2.5% retweets, 1.7% replies). Other countries get only 2% of retweets or replies although they author over 13% of tweets mentioning MPs.

As expected, a large fraction of mentions ($\approx 75\%$) come from the UK, but MPs show selectivity, with over $\approx 90\%$ of their retweets and replies being made to UK-based Twitter users. This suggests that *Twitter is serving as a way for MPs to keep in touch with the UK electorate*. MPs are even more responsive to Tweets from within their constituency (identified as mentioned in the Dataset section). As shown in parenthesis in Table 3.2, among incoming (⇐) Tweets from within the UK that mention MPs, only about 28.8% come from within the constituency region. However, MPs' outgoing (⇒) tweets prioritise interactions

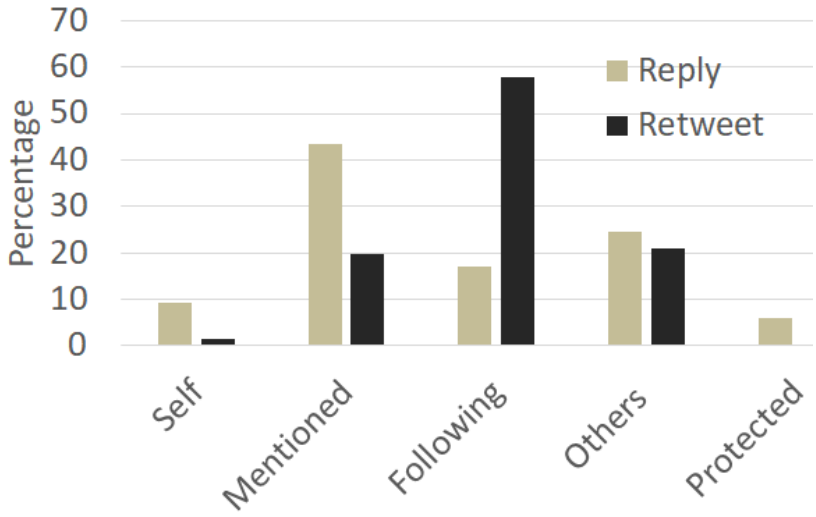


Fig. 3.4 MP responses (retweets and replies) put into mutually exclusive categories. A 'Self' response is a retweet or reply to MP's own tweet or reply. 'Mentioned' is a response to a tweet mentioning the MP. 'Following' is a response to a tweet which does not mention the MP but appears on their timeline because the MP follows the person. 'Protected' tweets and replies are not available to analyse. 'Others' comprises the remainder of retweets and replies.

with such local tweets: 56.7% of retweets, and nearly 60% of replies are focused within the constituency region represented by the MP[13].

I then compare the responses – replies and retweets – sent to different categories of people. Focusing first on the replies, Figure 3.4 shows that ≈43% of replies are to tweets that mention the MP directly; thus MPs are using their replies to engage directly in conversation with those that mention them on Twitter. In contrast to replies, most (57.8%) of the retweets are for those that the MP follows. In other words, MPs are retweeting other users even without the MP being mentioned. This is not surprising, since Tweets from those that MPs follow appear on MPs' timeline, and MPs may retweet what they find interesting. However, a disproportionate number of retweets are *tweets of other MPs, and in particular, MPs from the same party*: on average, other MPs constitute 7.4% of the following numbers of an MP. However, nearly 17% of all retweet actions are made on Tweets of other MPs. A further 6% of retweets are for posts made by their party's official Twitter handle. Nearly 96% of

[13]Note that this analysis only includes the 78% of Tweets for which I am able to extract a valid geographic location of the Twitter profile with whom an MP is corresponding. Also, I conservatively remove 60 London-based MPs from consideration because most MPs interact with journalists, lobbyists etc., who tend to be based around London.

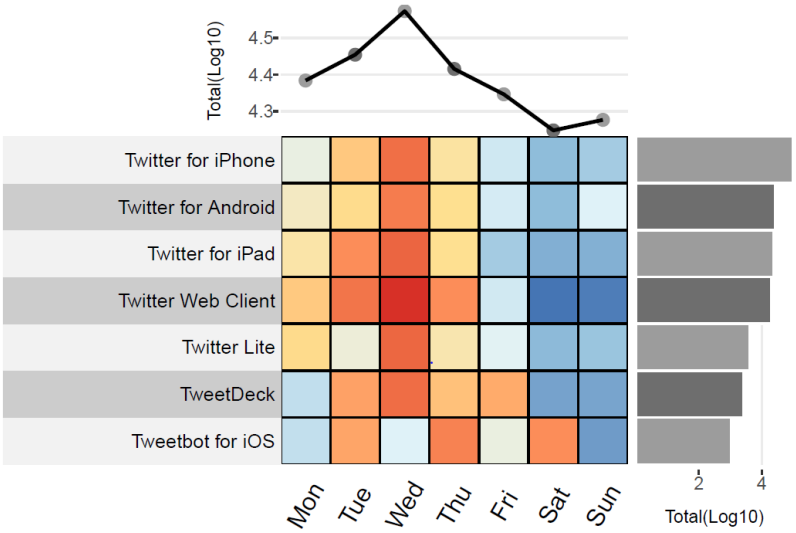


Fig. 3.5 MPs' outgoing (⇒) activity by day of week and Twitter client used. The 'row bar' on right side and 'col line' on top represents the counts of total data by Twitter client and day of week respectively. Dark blue represents the lowest activity and dark red the highest. Lowest activity is found trivially on Saturdays and Sundays and the highest activity is on Wednesdays, corresponding to Prime Minister's Questions. Android, iPhone and iPad clients are the most popular.

the MP-MP retweets are for MPs from the same party. Thus, it appears that *MPs are using retweets as political marketing, to boost their party's message* (termed as party maintenance by [Sta08]).

Figure. 3.4 also shows that a small but significant minority of replies (9.4%) are from the MP to themselves. This turns out to mostly be Tweetstorms – a single post which has been split into series of related tweets (posted in quick succession) because of Twitter's character limit. On Nov 7 2017, close to the midpoint of my data collection period (Oct 1–Nov 29), Twitter did expand the character limit from 140 to 280, but this hardly affected the volume of self-replies: Prior to Nov 7, there was an average of 36.02 self-replies per day from all MPs, and after this date, the average was 34.17 per day. Thus, it appears that in many cases, MPs need a larger text limit than 280 characters to discuss substantive topics.

3.5.2 Help from Staff

The previous subsection identified selective responses and prioritisation of constituents as one way for MPs to cope with the load of engaging on Twitter. As an alternate or complementary strategy, MPs may also employ staff designated as Communications Officer or Senior Communications Officer. Permitted (non-party political) activities of such staff include establishing a social media presence in the constituency, publicising surgeries, following up on social media queries and comments, publicising the MP's parliamentary duties on social media and proactive and reactive communications with all media [IPS18].

I cannot determine with certainty which tweeting instances originate from MPs and which from their staff, but I can find evidence suggestive of this. For instance, if multiple people are managing an account, it has the potential to be detected as a bot by the Botometer tool [DVF$^+$16, VFD$^+$17]. Nearly 8% MPs (45 of the 559) in my data are also detected as bots by this API[14]. I can also look at the Twitter client(s) used, as identified by the 'statusSource' field of Tweets collected from the MPs' handles. A total of 47 different sources are used by the 559 MPs, ranging from the Twitter Web Client and TweetDeck, to Twitter for iPhone or Android. Figure 3.5 shows that posting activity is mainly through the Web, whereas replies and retweets happen through personal devices such as iphone or android smartphones. Web clients can potentially come from multiple computers belonging to different staff. I also find that MP Twitter handles which use iPhone do not tend to also use android, and vice versa. Furthermore, the Web Clients are active mainly during weekdays. These patterns are suggestive of the MPs themselves, or one selected member of their staff handling the responses (replies and retweets), with the possibility of multiple staff being delegated the duty of posting new tweets, which may consist of advertising the MPs' activities; sharing videos and transcripts of their speeches etc. Note that the highest activity for the Web Client is on Wednesday, corresponding to Prime Minister's Questions, which, as mentioned before, is a highly advertised and popular activity. I also find that for over half the MPs, more than 80%

[14] https://osome.iuni.iu.edu/

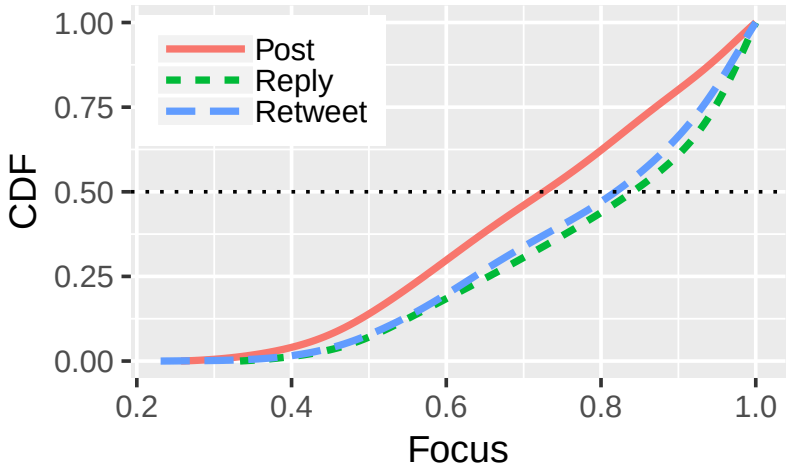


Fig. 3.6 Cumulative Distribution Function (CDF) of the fraction of an MP's activity which is captured by their most commonly used source. For over half the MPs, more than 70% of their original tweets (posts), and 80–90% of their retweets and replies come from one source, which could imply one person managing their handle.

of their replies and retweets come from one source (Figure 3.6), which is further indicative of one person managing their Twitter presence.

These observations can potentially be explained by the rule that MPs may not claim for party political and campaigning activities [IPS18].

Some activities identified above, such as retweeting their party position, may not be allowable due to this rule, and would therefore need to be undertaken by the MP rather than their staff. This hypothesis is in alignment with finding in Figure 3.6 that posts (original Tweets by MPs) are more likely to come from multiple sources than retweets – recall that posts tend to advertise MPs' parliamentary duties such as speeches and remarks made in the House of Commons, whereas retweets tend to amplify messages of other party members or the official party handle. MPs are also more likely to spend their limited budgets on communications staff only if certain conditions are met, for example MPs who are more junior and need to advertise themselves, or are in marginal seats [AU18].

MP / UK Users	Con.	Lab.	SNP	Lib. Dem.	DUP
Con.	0.70	0.26	0.01	0.02	0.00
Lab.	0.42	0.54	0.02	0.02	0.00
SNP	0.35	0.20	0.43	0.01	0.00
Lib. Dem.	0.45	0.25	0.01	0.28	0.00
DUP	0.26	0.15	0.02	0.01	0.56

Fig. 3.7 **Cross-party conversations in** $\Leftarrow$ Fraction of mentions from citizens who support (follow) MPs from one party to MPs of other parties. Each row adds up to 1, including mentions from citizens to MPs of their own party.

3.6 Tone of Political Discussion

Finally, I move to the third research question, and inquire about the nature and tone of the Twitter conversations. I am motivated to understand whether this platform, which appears to have gone mainstream in just five years since the first studies, and is being widely used by nearly all MPs, is contributing positively to the political debate.

This is an important question to answer, as various events such as Brexit have led to a highly charged and polarised political atmosphere in the UK, and both scholars and broadsheet newspapers have argued that the "middle has fallen out" of UK Politics [JCG17, SB17, Whe17, Bol16, GBG+19]. Note that although Brexit is polarising, it is not polarised among party lines. MPs from different parties have different stance on Brexit[15]. There is also wide concern that political discussion on Twitter involves aggressive and "trashy" language [Hin16, CRF+11].

Given the large scale of the data, I take a broad-brush approach, and focus on understanding whether there is cross-party political discussion between MPs and citizens, and on the tone and sentiments of the discussion as discoverable by tools such as LIWC 2015 [PBJB15].

UK Users / MP	Con.	Lab.	SNP	Lib. Dem.	DUP
Con.	0.66	0.30	0.03	0.01	0.00
Lab.	0.12	0.86	0.01	0.01	0.00
SNP	0.11	0.20	0.68	0.00	0.00
Lib. Dem.	0.29	0.38	0.06	0.27	0.00
DUP	0.22	0.21	0.01	0.00	0.56

Fig. 3.8 **Cross-party conversations in** $\Rightarrow$ Fraction of MPs from one party replying to citizens who follow (support) MPs of other parties. Each row adds up to 1, including replies from MPs to citizens of their own party.

3.6.1 Cross-Party Political Conversations

To quantify polarisation, I divide users based on the party they support (using the method specified in the dataset section), and ask the extent to which they interact with MPs of other parties. My focus on communications between people in power (MPs) and ordinary citizens distinguishes us from previous work that looked at how ordinary users have polarised [YB10, GW17, BHMDW15, CRF$^+$11]. Additionally, the UK is a multi-party system, which provides an interesting differentiating dimension to prior work, which has typically looked at a two-way polarisation, focusing on two sides of a conflict [YB10, BHMDW15], or on two-party systems like the USA [CRF$^+$11, GW17].

In Figure 3.7, I examine Tweets from UK users that mention MPs, and find that regardless of the party they support, there is a lot of cross-party talk. Specifically, I focus on the top five parties in terms of MP numbers, and find that supporters of all parties tend to tweet mentioning MPs of the Conservative party, which is currently in power. Exploration with a simple LDA (Latent Dirichlet Allocation) topic modelling[16] suggests that citizens are interested in mostly three kinds topics such as the budget, Brexit, and resignations of ministers (due to scandals during the period of the data collection). LDA is a probabilistic modelling [Ble12] and helps in identifying the latent topics of a corpora using co-occurring terms across tweets. Some of the most salient terms that occur in the examined tweets are

[15] https://www.theguardian.com/politics/ng-interactive/2016/feb/23/how-will-your-mp-vote-in-the-eu-referendum
[16] https://knowledger.rbind.io/post/topic-modeling-using-r/

Twitter, UK, Theresa, Brexit, People and so on. All of these topics have a natural focus on ministers and MPs of the governing party, which helps explain the surprising amount of cross-party mentions. Conservative supporters have the largest proportion of within-party mentions (69.8%). This suggests that users' *Tweets are directed at topical and current issues, and people involved in those issues, rather than the MPs they follow and the party they support, indicating a healthy attitude of engagement beyond the "echo chamber" of people who have similar views* in online conversations.

An extreme example of cross-party conversation is the case of the Liberal Democrats (Lib Dems), whose supporters have more mentions of Conservative MPs than the 12 sitting MPs whose Twitter accounts they follow (Fig. 3.7). In turn, Lib Dem MPs talk more to Labour supporters (who are more numerous) than those that follow them (Figure 3.8). A similar discrepancy between the interests of the MPs and its party supporters is observed with MPs of the Scottish National Party (SNP), whose replies have a greater proportion of replies to Labour supporters than conservative supporters, whereas ordinary citizens who follow SNP MPs talk more with conservative MPs than to Labour MPs. SNP and Lib Dems are ideologically closer to Labour than the Conservative Party[17]. I therefore conjecture that the discrepancy may be caused by MPs replying to those of a similar ideology as them, whereas citizens, who take a more questioning attitude (see next section), are engaging directly with the opposing view of the Conservatives.

3.6.2 Language and Sentiments

Given the surprising result of substantial cross-party talk between MPs and users, it is natural to ask what the tone of the conversation is between MPs and citizens – i.e., whether the discussion between MPs and citizens is a civilised discussion or an aggressive slanging match as in recent political campaigns [Hin16]. To measure this, I use LIWC[18] to summarise the language of the citizens mentioning MPs ($\Leftarrow$) on the one hand, and MPs replies to citizens'

[17] https://yougov.co.uk/news/2014/07/23/britains-changing-political-spectrum/, https://www.politicalcompass.org/uk2017

[18] http://liwc.wpengine.com/compare-dictionaries/

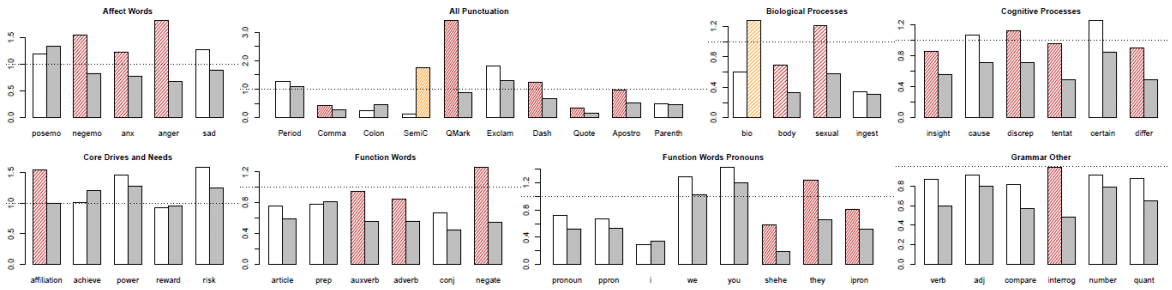


Fig. 3.9 LIWC scores for variables or categories reported by LIWC. Scores are normalised by dividing by the base rate scores expected by LIWC; if score is greater (resp. less) than 1, as marked by the horizontal dotted line, LIWC score is more (resp. less) than base rate. Corpus of $\Leftarrow$ tweets mentioning MP names is marked white; $\Rightarrow$ tweets from MPs to citizens are marked grey. A category is marked in red (resp. orange) if the score for $\Leftarrow$ (resp. $\Rightarrow$) is 50% more than for $\Rightarrow$ (resp. $\Leftarrow$) corpus.

Tweets ($\Rightarrow$) on the other. I look for differences and similarities in language usage between these two categories, to understand the tone of the discourse between MPs and citizens.

LIWC provides 94 dimensions along which to measure different aspects of language use [PBJB15, TP10]. For each dimension, it also gives the base rates of word counts to be expected in normal usage. Figure 3.9 shows the LIWC scores obtained for each dimension of language use, for both $\Leftarrow$ and $\Rightarrow$, normalised by the expected base rates of usage of that dimension.

I focus on the LIWC categories which see higher than base rates of usage or where there is substantial ($> 50\%$) difference between $\Leftarrow$ and $\Rightarrow$. Using this approach, I can make the following observations from Figure 3.9:

1. Both MPs and citizens show more positive emotions than LIWC base rates, which could be suggestive of a respectful or appreciative discussion.

2. However, citizens' incoming ($\Leftarrow$) Tweets also express more than the base rate of negative emotion, anxiety and anger, suggesting more conflict in some conversations. The citizens' Tweets also raise a lot of questions, heavily using interrogatives, and question marks in their language. This could potentially be related to scandals and related resignations during the period of study.

3. Unusually for political conversations, there is a large amount of sexual, sex-related and swear words, owing to the Westminster sex scandal which erupted in the wake of the #MeToo movement, and led to the resignation of several ministers and MPs during the period of the study[19]. Table 3.3 shows that MPs appear to take into consideration the language of a mention in deciding whether to reply back. Incoming (⇐) mentions from citizens which exhibit anger or use swear words are less likely to elicit a reply.

4. Perhaps because of this scandal, citizens' incoming (⇐) tweets towards MPs have a higher than base rate of "moralising" language, using words such as 'should', 'would' (marked as discrep), and 'always', 'never' (marked as certain).

Table 3.3 shows examples of tweets with expressing various dimensions[20] of tone which are less likely to get a response from MPs. While some categories like swear, question marks, anger, filler words and other non-fluencies (e.g., "err", "um", I mean, you know, etc...) so on indicate that MPs are responding less to such tone, other tones like shehe, leisure and male are interesting cases and require deeper examination of the tweets. For example, shehe signifies more use of third-person speech which could be a pattern of how users mention MPs. Leisure can relate to funny comments made to mock MPs and male reference could be more dominant because there are more male MPs in the dataset.

This broad-brush approach is intended only to provide a flavour of the tone of discussion. It appears to indicate that in the study period, which was rich in scandals that affected multiple MPs and ministers, citizens are using Twitter as a platform to freely and directly question their representatives and express negative emotions, anxiety and anger, as they are entitled to. However, they also show higher than base rates of positive affect and appreciation where warranted. In return, MPs appear to exercise restraint, using higher than base rates of positive language, and avoiding using or responding to negative language (which could escalate conflict). I conjecture that the public nature of Twitter leads to MPs being conscious of the effect of their words on their image and public perception, providing a platform for civilised

[19]Wikipedia provides the most up to date account at https://en.wikipedia.org/wiki/2017_Westminster_sexual_scandals.

[20]LIWC 2015 Development Manual: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

Dimension (Description)	Difference	Example Tweet
Swear (Informal aggressive tone)	90.5%	@MP c**t.
QMark (Punctuation '?')	79.2%	@MP WTF?
Filler (Informal unprepared speech)	78.4%	@MP Blah blah blah.
Exclam (Punctuation '!')	72.2%	THANK YOU @MP !!!!!!!!
Negate (Negations with words)	66.4%	No @MP. Not surprised...
Nonflu (Nonfluencies: er, hm, umm)	63.2%	@MP Well....
Shehe (Mention to third-person)	63.2%	She's @MP constituent isn't she?
Anger (Feeling of annoyance)	58.2%	@MP F**king s**t stirrer
Leisure (Activities: cook, chat)	58.1%	@MP Nice joke.
Male (Reference to gender)	56.6%	Ignore him. He's "Loyal" @MP.

Table 3.3 LIWC categories which make it less likely that MPs respond back to citizens. This table shows the percentage difference of some LIWC dimensions that corresponds to the decreased likelihood of MPs in making responses to incoming ($\Leftarrow$) mentions if these LIWC categories are present. Some example tweets are also listed which appear in the dataset. Note that for extreme abusive words shown above I have replaced some character by *. Also, MPs Twitter handles are replaced by @MP.

discourse. I note that this kind of behaviour may be partly due to the focus on interactions between MPs and citizens. Previous studies in the UK context have found that MPs have indulged in attacks on other politicians, especially in election contexts [GBHVH13].

3.7 Case Study 1: A Possible Future of Online Twitter Engagement

The first case study understands a novel way in which the immediacy of Twitter was used to improve democracy by allowing citizens a part in creating an Act of Parliament, and discuss how it speaks to three research questions:

Individual MPs in the UK Parliament are able to submit Bills (also known as draft legislation). These are known as Private Members' Bills. Priority is given to Government-sponsored Bills, so to ensure that a proportion of Private Members' Bills have a chance to become law, there is a ballot of MPs each year to assign priority for the limited amount of debating time available. However, even Bills coming high up in the ballot are unlikely to be passed unless they have the tacit or explicit support of the Government.

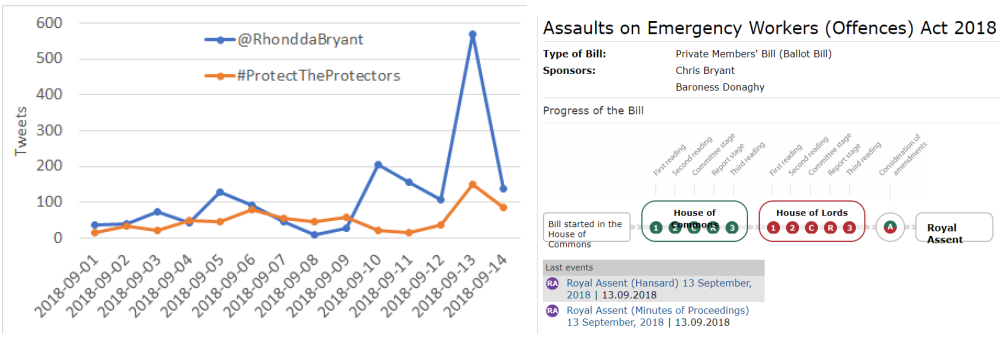


Fig. 3.10 Left:#ProtectTheProtector tweets burst on the day of Bill getting passed in House of Commons. Right: Status of the Bill as on 15^{th} Sep 2018.

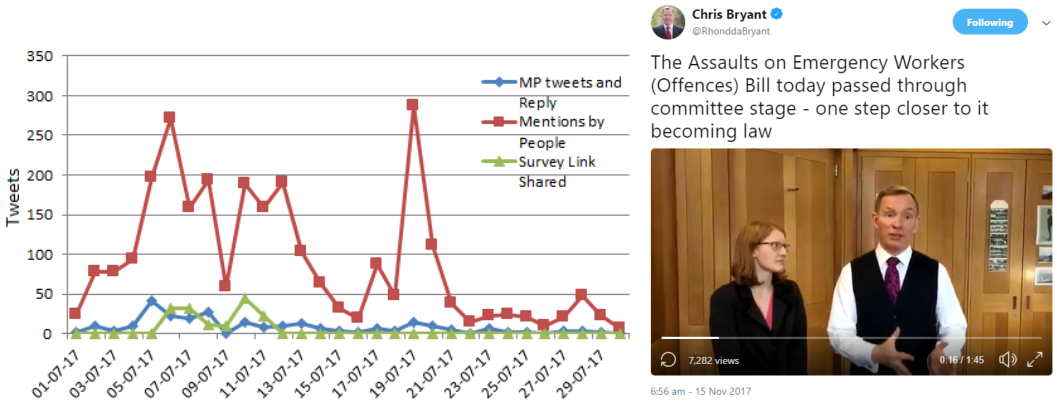


Fig. 3.11 Left: Survey tweet links, MP Chris Bryant ⇒ and ⇐ Right: MP Chris Bryant's thanking video after his proposed bill was passed the Committee stage.

Chris Bryant, a Labour back bench MP, came top in the ballot for the 2017-19 session, and therefore was eligible to propose a Bill. However, as a member of the Opposition Party who is also not among the prominent "front bench" MPs, his bill would have faced an uphill task. Bryant launched a consultation on Twitter in July 2017 (before the study period), putting forward six possible Bills and asking Twitter users to choose their favourite through an online survey. 45,000 people participated[21], and the winner of the poll was a proposal to provide additional legal protection to emergency service workers, as a result of reports of assaults by members of the public during emergency call-outs. Bryant introduced the Assaults on Emergency Workers (Offences) Bill 2017-19. This bill was greatly strengthened

[21]https://www.youtube.com/watch?v=RqkSSQ3bAaA

by the evidence of public support, and was one of the few Private Members' Bills supported by the Government. Tweets during the passage of the Bill through Parliament used the hashtag #ProtectTheProtectors, and received a high level of engagement. On 13 September 2018, the Bill received Royal Assent, the final stage on the way to becoming a law. It has now been signed into law as the Assaults on Emergency Workers (Offences) Act 2018.

This innovative approach, and the effective use of Twitter, surveys and hashtags has enabled the public to follow the progress of the Bill throughout its timeline and participate by providing direct comments, creating an experience closer to direct democracy [Col05a]. It has also acted as a means of garnering publicity for Bryant. This can relate to the research questions: RQ-1 and RQ-2 seek to understand how much load is incurred by the MPs, and how they manage this load. Clearly, with 45,000 responses to the initial survey, this was a huge effort. The use of a survey tool was critical to manage this huge load and summarise their response. However, the MP also struggled to cope: Analysis of Tweets during the survey/poll in July 2017 reveals a typical pattern of activity during a focus period (Figure. 3.11). Twitter users mentioned the MP (in red), often to make suggestions or enter into dialogue, but as shown by the blue line, he was unable to respond to many. This showcases both the potential for direct and participatory online engagement via Twitter but also the drawbacks, if excessive activity makes a personal response impracticable.

To study RQ-3 on the nature and tone of the discussion, I focus on the final day of high activity during the passage of the Bill. Figure 3.10 shows that the MP gained more than 500 mentions on just one day (Sep 13), when the Royal Assent was obtained[22]. With a high number of mentions like this on a single day, it is hard to respond to each mention; reiterating again that although the process innovatively unlocked the participatory potential of Twitter, the burden of response during such direct engagement remains an issue (RQ1). To manage the load, the MP did a 'thank you' post as a collective response to all (RQ2). As an event where high attention was anticipated, the messages were mostly appreciative and complementary to the MP[23]. Although as expected the majority of congratulatory Tweets

[22]https://services.parliament.uk/bills/2017-19/assaultsonemergencyworkersoffences.html

[23]The volume of the whole conversation from July 2017–Sep 2018 permits only a cursory examination, but also seems to incorporate a mostly civil and respectful tone; with suggestions and requests for changes,

were from Labour supporters, it is remarkable that close to 28% of tweets come from non-Labour supporters, showing the broad multi-partisan support for the Bill, offering hope for constructive participatory democracy through the innovative use of digital tools like Twitter.

3.8 Case study 2: Hate Speech in UK Political Discourse

In recent years, online presence has become an essential component of modern democracies [Jun16, GDFMGM18, PP19, HBKH20, GMBG10]. Although this increasing digital reach is bringing more people into politics [Vac13], increasing online activity in the political and other spheres has led to concerns about online hate. Evidence suggests that online hate is a grave and growing problem. Not only does it cause short-term frustration, anger or fear to its direct and indirect addressees, but it may also have long term implications on victims' mental health or marginalise them and dissuade them from actively participating in public discourse [VBM21].

In this second case study, I shed light on this important issue using the dataset in this study (containing 2.5 Million tweets from 293k users as well as 579 MPs of which nearly 1% is characterised as hate (Section §3.3)). Unlike most previous efforts [GBR+20], the data captures entire threads of conversation between Twitter handles of MPs and citizens in order to provide context for content that may be flagged as 'hate'. Having set up the raw dataset together with associated metadata about the MPs and citizens as well as hate labels for each tweet, I am now in a position to more closely examine the prevalence of hate speech in this nationally important conversation between citizens and their elected representatives. Specifically, I ask *who is targeted* by hate speech in this conversation (generalisation RQ1 and RQ2) – i.e., whether there are specific parts of the dataset where I may find more hate speech than in other parts of the dataset. To this end, I slice and dice the data in different ways and establish how many hate-labelled tweets I find in each cluster of tweets formed (as presented around RQ3).

inquiries as to why the bill is required when assault by itself is already considered a crime, as well as messages of encouragement.

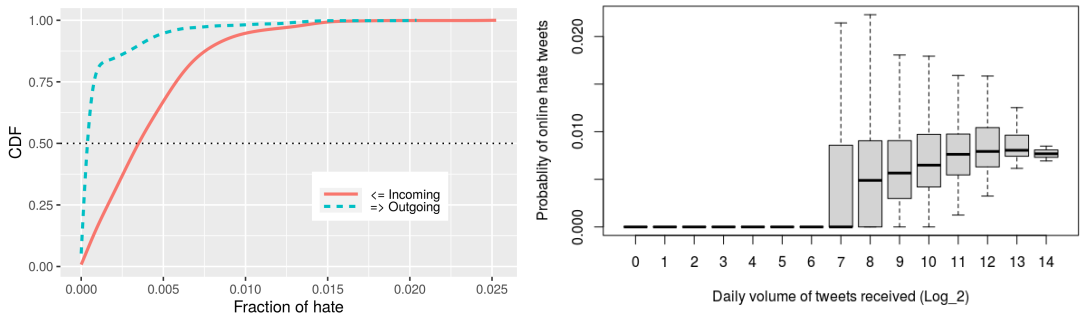


Fig. 3.12 Left: Fraction of hateful tweets per user ($<=$ *incoming*) towards MPs and per MP ($=>$ *outgoing*) towards non-MPs. Right: Probability of receiving hate as a function of the volume of tweets received that day by an MP.

Incoming and outgoing hate: I divide the data into tweets authored by MPs and tweets authored by others but directed to (i.e., mentioning or replying to) MPs. Figure 3.12 (Left) shows the cumulative distribution of the fraction of tweets with hate labels in tweets made by MPs vs. tweets directed at them. This clearly demonstrates that there is much more hate directed at MPs although 67 tweets made *by* MPs do get classified as hate speech. Many of these contain strongly expressed opinions which may perhaps not be strictly considered 'hateful' by manual annotators (e.g., "What a stupid tweet. So you would prefer Daesh to still control It? Even your idol Putin does not want that") or use of violent or rude words in a humorous context, which may confuse the hate speech classifiers ("@hugorifkind Off with your head!"). When MPs receive hate speech, it appears that there is a higher probability of receiving hate on days when they receive a high volume of mentions. Figure 3.12 (Right) shows that after a certain threshold number of mentions per day, the probability of some of those mentions containing hate speech rises dramatically. MPs tend to get a high amount of attention (mentions) when they are in the news for one reason or another. In some cases such attention is planned or anticipated (e.g., Prime Minister's Questions (PMQ) on Wednesdays is a highly anticipated event in Parliament and increases the volume of tweets towards the PM. Similarly the Chancellor of the Exchequer when he releases the budget). In other cases, an MP receives attention because of a controversy (e.g., International Development Secretary Priti Patel had to resign when it emerged she had held unofficial meetings with

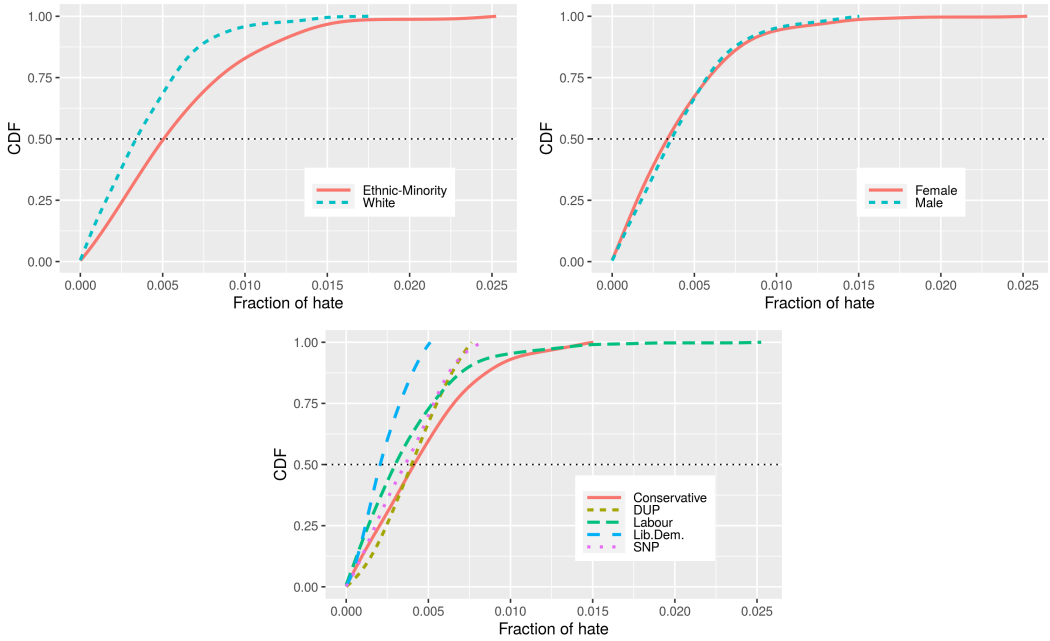


Fig. 3.13 Fraction of hate by demographic characteristic: Ethnicity (Top-Left), Gender (Top-Right) and MPs by political party (Bottom).

Israeli politicians and officials whilst on holiday in that country). In such cases, the increased amount of attention towards the MP appears to attract a higher than usual proportion of hate. This suggests that there may be some "pile on" harassment going on, whereby MPs receive more hate because of other hateful comments and mentions they are receiving.

Hate by MP demographics: To begin with, I break down the prevalence of hate by three demographic characteristics of MPs: ethnicity, gender and party. There is no official data that comprehensively records the self-defined ethnicity of MPs, but the British Future think tank compiles a list of ethnic minority MPs following each general election (cf. Section 3.3). The majority (92%) of the MPs were not members of ethnic minorities according to this list. Figure 3.13 (Top-Left) compares the distribution of hate speech received by white MPs to hate received by ethnic minority MPs. It can be seen that there is a statistically significant (KS stats: $D = 0.23, p < 0.02$) difference, with MPs from ethnic minorities receiving more hate than those from the white majority. A parliamentary enquiry has also expressed concern about openly available content such as Tweets which may stir up hatred against minorities [Par16]. The parliament in 2017 had 208 female MPs (32%), the highest number since women were

allowed to become MPs in 1918 [Par17]. However, online misogyny has been extensively documented [GVM$^+$21, EZ21, FS19]. Therefore, I next examine whether female MPs (of all ethnicities) get more hate than male MPs (of all ethnicities). Surprisingly, I find that (Figure 3.13 (Top-Right)) there is no statistically significant difference between male and female MPs. The next natural division to examine is along party lines. As mentioned previously, the Parliament in 2017 had MPs from a number of parties, with the Conservative Party being the majority party that formed the government. Figure 3.13 (Bottom) shows the distribution of the proportion of hate tweets received by MPs of each party. It can clearly be seen that MPs of the governing Conservative Party receive much more hate than MPs of other parties. Figure 3.13 (Bottom) breaks this down to Conservative MPs who hold a formal position in the government vs. all other MPs – MPs with ministerial positions tend to get more hate than so-called 'back bench' MPs who do not have a ministerial portfolio.

Next, I study how supporters of one party may interact with MPs of their own and other parties. Figure 3.14 (Top-Left) shows how the supporters of each party (as computed in §3.3) distribute their MP mentions among parties. Note that mentions labelled as hate are removed from the computation in Figure 3.14 (Top-Left) as they are taken up in Figure 3.14 (Top-Right). As expected, the highest proportion of (non-hate) mentions are towards MPs of the same party as the supporters. Each row in the left and middle figures sum to nearly 100%, except where there were mentions to the three other parties (not included in the figure as there are very few mentions). From Figure 3.14 (Top-Left), is interesting to observe that apart from the two parties with the most number of MPs (i.e., apart from Conservative and Labour Parties), the fraction of mentions to their own party is less than 50%. i.e., supporters of smaller parties talk *more* to MPs of all other parties collectively, than to MPs of their own parties. In large part, this appears to be because supporters of all parties tend to mention MPs of the Conservative Party, which is currently in power. Labour, which is the largest party in Opposition and forms a 'shadow cabinet', also receives a fair number of cross-party mentions. Figure 3.14 (Top-Right) shows the distribution of hate speech within and across party lines. As expected, here the roles are reversed with most of the hate speech going to MPs of other parties. However, again I observe that the party in Government, the

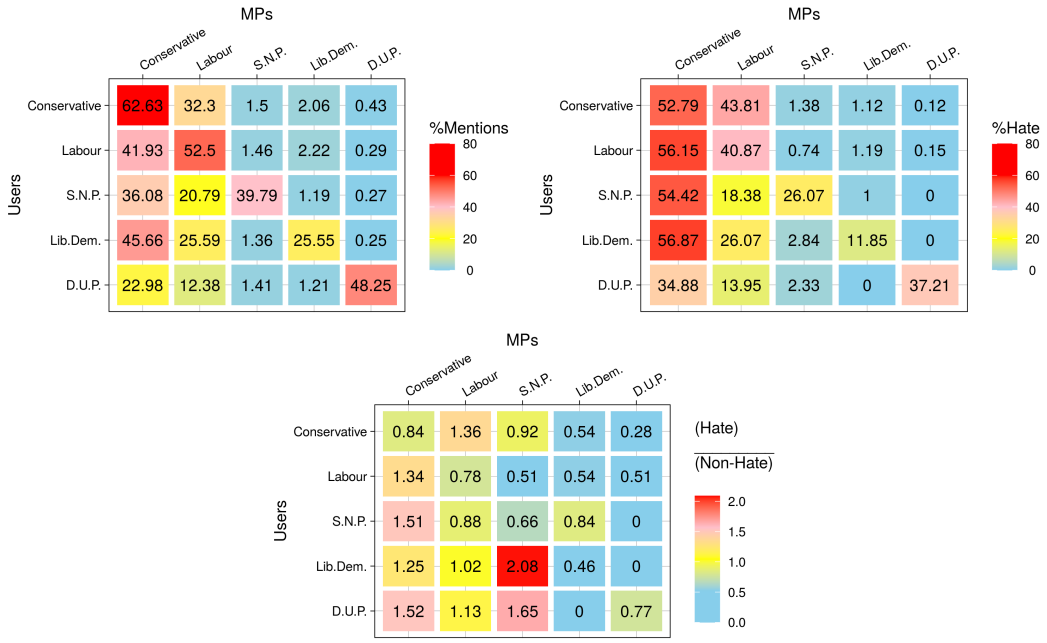


Fig. 3.14 Percentage of Cross-Party and Within Party Mentions (Top-Left) and Hate (Top-Right). Rows add up to nearly 100% in the Left and middle matrices. Bottom: Ratio of the percentage of hate (from the top-right matrix) to mentions (top-left matrix).

Conservatives, get a large fraction of hate, including from their own supporters. The *absolute* number of hate-labelled tweets comprises less than 1–2% of all mentions. However, it is interesting to see how the amount of hate speech between each party pair varies in relation to the volume of mentions between the same pair of parties. Figure 3.14 (Bottom) computes this as the ratio between the percentage of hate to the percentage of mentions (i.e., by dividing each entry in the matrix on the left by the corresponding entry in the middle matrix). I see here that within each party (i.e., along the diagonal of the matrix), the ratio is less than 1.0. In other words, there is a smaller proportion of hate speech in comparison with the volume of within-party mentions. In contrast, the party in power, Conservative Party, receives a higher proportion of hate than non-hate mentions across the board, from supporters of all other parties.

3.9 Discussion

As early as 1774, the political philosopher and MP Edmund Burke stressed the importance of understanding the views of constituents [Bur74]. Despite this early recognition of its need, active engagement with constituents outside the election period was rare until the mid-twentieth century [AU18]. However, it is now seen as a necessity by MPs, and this is being increasingly facilitated through digital means.

My research agenda involves understanding the usage of Twitter as a new form of continuous citizen engagement. In this work, I identified Twitter as an interactive platform which seems to have become part of mainstream usage, used by nearly all MPs, and with a high volume of activity. I investigated the dynamics of the load imposed by the increasing volumes of Twitter activity and the consequent attention directed towards MPs. I showed that attention can be highly focused, with a large proportion of the total activity directed at an MP occurring during short focus periods of 3–5 days. MPs use selective replies and prioritisation of local or constituents' concerns as a way to manage this high attention load. They use their Twitter presence strategically, balancing their role as party representatives with the role of hearing and responding to their citizens' needs. I also find that Twitter presents possibilities for immediate and direct discussion, leading to new possibilities for cross-party discussions on a level playing field and therefore holds promise for bridging, or at least initiating conversations, across the political divide.

I then ask whether certain kinds of MPs are targeted more than others. I find evidence that MPs from ethnic minority backgrounds receive more hate than MPs from white backgrounds. I also find that male and female MPs are targeted equally by hate speech[24]. However, I find that MPs from the governing Conservative Party, especially those with a position in Government (e.g., as a Cabinet minister) receive more hate than other MPs. I also find that hate comes from across the party lines, with supporters of one party attacking MPs from other parties.

[24]Note that to understand the type of hate, a deeper investigation of the content will be required. For example, cases of misogyny [Dug17, GVM+21]

My findings have important implications from a legal and policy perspective. They provide evidence-based support for the UK Law Commission's proposals for law reform to capture coordinated and non-coordinated "pile on" harassment [Com20a]. The fact that MPs having an ethnic minority background receive more hate on social media platforms should be taken into serious consideration by policy-makers. Further, with regard to the recent EU Commission proposal to regulate content moderation, my work provides an example of the kind of measurements that can be used as an element of social-media platforms' annual reports, should the latter become part of the forthcoming Digital Services Act [B$^+$21]. This work is a small step in combating the problem of hate speech in the national discourse, which many researchers see as a potential threat to social order, threatening social peace and cohesion [WBJ$^+$20].

CHAPTER 4

ECOSYSTEM OF POLITICAL INFORMATION AND DIGITAL MASS MEDIA PLATFORMS

If you are not paying for it, then you are the product.

Andrew Lewis (aka blue_beetle)

4.1 Introduction

With most political discussion happening in cyber-space, political parties and publishers are now heavily focused on providing user audience their news content via online means like websites, social median and Wikipedia (an online encyclopedia). Moreover, in the era of mass Web monitoring, users are being tracked and their behavioural data collected and used, typically for *personalisation*. In fact, a deeper examination of the online political ecosystem reveals a set of complex techniques such as ads trackers, synchronization of cookies across websites and fingerprinting of user devices [PKM19, OMDC14, AEE+14, NKJ+13, PDP+17], in order to perfect the user profiles for useful personalisation. Indeed, a highly precise profile allows ad-platforms to effectively match ads with target audiences. However, the side effect of these highly sophisticated profiling techniques is the undermining of user's privacy and emerging polarisation in political information which is available in digital mass media platforms. Without any consent, a profiled user is unknowingly disclosing her interests and (dis)likes to the benefit of mass media platforms.

Recently, such user profiling techniques have also been used *for targeted opinion shaping* such as Cambridge Analytica's U.S. presidential campaigns for Donald Trump [PL18] and Ted Cruz [SS18], UK-based campaigns for the Brexit referendum [GC18], and Russia's social media operations during the 2016 U.S. presidential election [Ber18]. All these campaigns have one common motif: use of partisan disinformation campaigns. In fact, a slew of websites hosting such content have emerged since the start of Trump's campaign [BJBS18]. These pages allow fake and partisan news to be shared unchecked on the social media, garnering ever increasing partisan audiences. Some examples are Infowars, Breitbart and Fox News, for the right, and Occupy Democrats, Bipartisan Report and MSNBC for the left.

At such a juncture, this research is compelled to ask: *Do some hyper-partisan news websites (HPWs) and other online political content sources– which have been shown to have highly selective audiences – exhibit any particular differential behaviour when it comes to delivering information to their online visitors?* i.e., do websites on the left (right), differentially track and personalise the left (right)-leaning visitors more than the right (left) leaning ones?

To answer this multi-faceted question, I first establish a methodology for understanding how HPWs and their embedded third parties track different user demographics. I create 9 carefully crafted personas representing different genders and age groups. I load browsers with these personas and visit a list of 556 verified HPWs [BJBS18], to observe differences in the way these personas are being tracked via different types of cookies placed by HPWs and the third party ad-ecosystem. I also offer a complementary study (as a case study, c.f. §4.7) which investigates the presence of polarisation in Wikipedia profiles of politicians, a widely-used source of political information. The content which is available to readers via politicians pages is created over-time by Wikipedia volunteers. I collect and characterise these pages based on their spatio-temporal patterns and a new form of polarisation which is based on news citations that exist in these pages.

With this chapter, I make the following main contributions:

1. I design the first to my knowledge methodology to detect possible unbalanced tracking performed from HPWs and their collaborating third parties of users that belong to different demographics.

2. I extend an existing crawling tool with my methodology, and open source my framework[1] in order to enable fellow researchers, policy makers or even end-users to audit websites on how they personalize tracking technology based on the visitor's web profile.

3. The results in this chapter show that in HPWs , *advertisers set the majority of cookies* on users' browsers. *These advertisers are among the top in the overall ad-ecosystem*, and are over-represented on HPWs compared to their presence on the general web. In fact, some of the top trackers are twice as prevalent on HPWs as on the general web. Also, having an established persona from a particular demographic (with cookies obtained from visiting stereotypical websites for users of that demographic) results in *up to 15% more cookies* stored than for a baseline with no set persona. Furthermore, popular or highly-ranked HPWs track users more intensely than lower-ranked websites. More importantly, my results show that right-leaning websites, in general, track users with *up to 25% more cookies* than left-leaning websites.

4. Finally, I present two case studies to understand personalisation and polarisation while engaging with news broadcasting. In the first case (c.f. §4.7), I search for large-scale patterns by co-clustering both the personas and websites visited, using non-negative matrix factorization [LS99, LGG18, ZYCG07]. In the second case (c.f. §4.8), I focus on Wikipedia profiles of Politicians. Wikipedia is a platform where thousands of volunteers revise and add content constantly [INPG10]. Interestingly, I find that engagement with Politicians Wikipedia articles is synchronized with election periods, and that a mild form of polarisation exists in editors' preferences and article references.

Next I provide background to some key concepts and then I discuss implications of research work for online users' privacy.

[1] Data and code available from http://tiny.cc/partisan-tracking

4.2 Background of Third Party Tracking on the Web

Websites nowadays consist of several components that may originate from many different domains other than the one the user visited. These components provide functionality like widgets, analytics, targeted ads, and recommendations. In order to provide as much personalized content as possible, these third parties keep track of user's personal information (e.g., geolocation, browsing device, gender) and preferences (e.g., purchases, searches, etc.) locally in a cookie placed in the user's browser, and in their servers' database using the same cookie ID.

Although several policies (e.g., Intelligent Tracking Prevention [Wil19] and Same Origin Policy [Wor10]) have been proposed to mitigate the privacy intrusion from such pervasive tracking, there are several sophisticated techniques that allow trackers to bypass such mechanism (e.g., Cookie Synchronization [PKM19, BARW16], Web Beacons [HRG01], Cross-Device Tracking [SIIK19], etc.) and re-identify a user across different websites and create detailed profiles about her preferences and interests (e.g., sexual preferences, political beliefs, etc.) [JM09]. Then, this data can be sold to anyone interested [The00], or handed over to agencies [AK13].

4.3 Background of Wikipedia in Politics

Information about politics on Wikipedia would appear to fulfil at least three of the functions of the communication media in 'ideal-type' democratic societies as described by McNair: informing citizens of what is happening around them; educating as to the meaning and significance of the 'facts'; and publicising the activities of governmental and political institutions [McN11].

Wikipedia is a major source of information providing a large variety of content online, trusted by readers from around the world. Wikipedia's political content is pretty extensive. For example, all 650 elected Members of the Parliament (MPs) in the United Kingdom (UK)

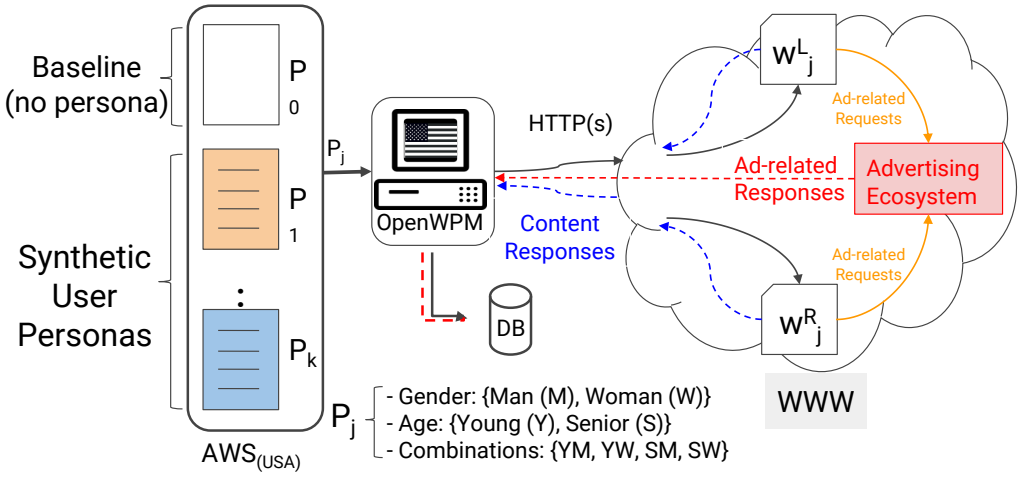


Fig. 4.1 Crawling methodology and framework for measuring tracking of different personas by hyper-partisan websites visited, and third parties embedded in them.

have a Wikipedia page or article[2] Such content is also widely reused by the broader web: due to the popularity of Wikipedia as a platform, the top results returned by search engines queried for MPs show links to their Wikipedia pages. Political actors may therefore regard Wikipedia as an attractive platform for advertising elected representatives and candidates by controlling the information flow or enhancing their image [GM18].

Certain editors proclaim their political leaning and form communities [NLK+13]. Polarised teams–those consisting of a balanced set of politically diverse editors–may create articles of higher quality than politically homogeneous teams [STDE19]. I discuss this dataset in detail in §4.5.3 and then findings in §4.8.

Table 4.1 Terms and notations used in the methodology.

Terms	Notations
General persona	P_j, where P_0 is baseline (i.e., does not have any profile or user history loaded),
Simple persona	$P_{1 \leq j \leq 4}$ are 1-feature personas,
Rich persona	$P_{5 \leq j \leq 8}$ are 2-feature, "rich history" personas
Hyper-partisan websites (Left)	W^{L_i} are left-leaning websites; $1 \leq i \leq 164$
Hyper-partisan websites (Right)	W^{R_i} are right-leaning websites; $165 \leq i \leq 556$
Tracking Domains (Left)	D^{L_i} are domains in all left-leaning websites; $1 \leq i \leq 164$
Tracking Domains (Right)	D^{R_i} are domains in all right-leaning websites; $165 \leq i \leq 556$
Crawl (SQLite Database)	$C_n^{P_j}$ represents crawl database file for persona P_j, $0 \leq j \leq 8$ and $1 \leq n \leq 5,15$ runs of the same setup depending if it is a simple persona or a rich persona.

4.4 Methodology

4.4.1 Overview of Crawling Methodology

A general illustration of the methodology and implemented framework is shown in Figure 4.1. To obtain consistent user behaviour that allows repeatable and systematic web measurements, the proposed framework leverages a methodology with carefully curated user personas that have a history built on Web traffic that could be expected from users of specific demographic groups. Using each of these defined personas, I then visit HPW , which have been carefully categorized (manually) as left- or right-leaning by domain experts (journalists and fact checkers at Buzzfeed News [SLTVSV17b, SLTVSV17a]). Then, I log and monitor first and third party (e.g., ad-ecosystem) tracking performed during these visits, for later analysis and comparison between personas and baseline or null users. I summarize various terms used in the paper in Table 4.1.

[2]I use the terms 'page' and 'article' interchangeably.

4.4.2 Incremental Persona Building

My goal is to visit different websites with the "persona" of a user from a particular demographic group, based on age or gender. I build personas by visiting top Alexa ranking websites for different demographics. The intuition is that the third party ecosystem that enables tracking and personalized targeting [CMC+15, Ade19, BHK16] builds up a persona or profile of a user based on the type of websites they visit. Thus, by visiting websites which are highly popular in a particular demographic, I establish a persona of that demographic. The personas created and used in this study are listed in Table 4.2 along with some examples of websites from each category. As also shown in [BJBS18], Alexa rankings websites for age based demographic have Youth between 18 years to 34 years and Seniors above 65 year. Previous works have inferred that there are certain demographic and geographical features which are highly associated with a person's partisan lean [BJBS18], including age, gender and location (whether from a right- or left-leaning US state). Therefore, I build the personas based on these characteristics, focusing on age and gender, and training from different locations in the USA.

I use `alexa.com` which gives the list of websites that are most popular with different demographics[3], as captured in October 2018. However, an important question is: *How many of the websites should be visited, before a user profile becomes "stable"?* To check this, I visit the top websites in each category in a random order[4], and count the number of distinct third party cookies obtained after each new visit. Figure 4.2 shows the number of third party cookies for personas with different demographics, after visiting the top websites for each category. I observe a large increase in the number of cookies when the first 2–3 websites are visited, and then the increase in absolute numbers of distinct cookies tapers off. Following the average number of 30 cookies found per website in a study of 1 million Alexa-ranked websites [EN16], I say that a persona has "matured" once it has reached ≥ 50 unique third party cookies. I observe that visiting 6–7 websites is sufficient for a persona to mature, except in the case of Youth, which does not receive more than 100 cookies (possibly due to

[3]Under https://www.alexa.com/topsites/category

[4]I have verified that the actual order of visit does not affect the results

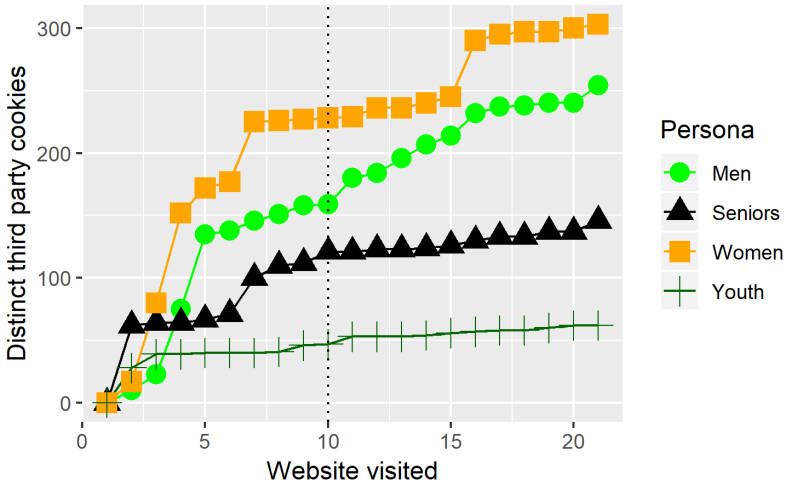


Fig. 4.2 Numbers of distinct third parties observed after visiting a set of top `alexa.com` websites for building persona per category (man, woman, young, senior).

laws such as COPPA [Com98] which safeguard against collecting personal information on very young users) even after visiting 20 websites. However, even in the case of the Youth demographic, I observe that the first 2–3 website visits already expose the user to the majority of distinct third party cookies, and by the time 6–7 websites are visited, well over 50% of the eventual total number of cookies have been collected.

Therefore, I conservatively assume a persona with a specific demographic being "stable" or "established" by visiting the top 10 websites associated with users of that demographic, and storing the third party cookies obtained. Also, visiting the same websites from different locations produced similar numbers of cookies. When visiting other websites (e.g., the HPWs in §4.5), I mimic a specific demographic by loading cookies obtained for that user group. Similar numbers of websites have been used in other studies to establish a profile of cookies. For example, in [LSS+15] they used the top 10 websites of each of 16 categories for analyzing ad content.

Note that I also use compound personas with more than one demographic characteristic (e.g., young man, or senior woman). To establish compound personas, I visit and store cookies from the top 10 websites from each demographic (e.g., 10 of the top websites for seniors and 10 for women collectively establish the "Senior Woman" persona), in a random order. I encode each persona with the initial letter of the feature incorporated in the persona.

Table 4.2 Examples of websites from each category of `alexa.com` for creating personas with specific demographics.

Demographic	Sample websites
Youth (Y)	(Student, Kidzworld).com, Voicesofyouth.org
Senior (S)	Aarp.org, (Medicare, Cms).gov
Woman (W)	(Cosmopolitan, Womansday, Sheknows).com
Man (M)	(Menshealth, Mensjournal, Esquire).com

For example 'SW' signifies a persona representing *Senior-Women*. Similarly a YM persona represents the persona of a Young-Man.

I visit all websites in a persona-specific list as a stateful crawl that stores the user history in a browser. After the visits are finished, I dump the browser state accumulated as an archive file per persona. This persona state corresponds to stored cookies, HTTP calls done, location, etc. Each archived file can then be used to bootstrap a browser with the said persona, ensuring my measurements per persona for each visit of a HPW are bootstrapped with the same browser persona state. Next, I discuss more details on automation and browser settings of the crawling approach.

4.5 Datasets

4.5.1 Crawling Engine of Framework

I make use of *OpenWPM* [EN20], a popular tool for measurements and automating web browsers, which is developed by Princeton University under the Web Transparency & Accountability Project [EN16]. Using OpenWPM, I create user personas and bootstrap parallel sessions of *Firefox* browsers to visit HPWs with each persona. In order to be able to compare browsing experience of a persona with a baseline, I load null personas (i.e., a user with no website visit history) to the browser and re-execute the website auditing. Also, using the same tool, I collect cookies set both by javascript calls (i.e., transactional cookies) and HTTP calls (regular cookies set in the browser, logged as Profile cookies). Transactional

cookies include all logs of cookies set in the browser, including the deleted instances of cookies. In my analysis, I consider both types and refer to them as *Cookies*.

I kept most of the default settings of OpenWPM's browser instance and updated a few for my use. The default settings include loading a Firefox browser (only this is supported by the tool), having always enabled third party (tp_cookie) tracking, keeping blockers like donottrack, disconnect and ghostery, https-everywhere, adblock-plus, ublock-origin and tracking protection as false, so that nothing is blocked. The updates include enabling the http_instrument which logs HTTP responses, requests and redirects, using the selenium headless browser to perform crawling and, setting js_instrument, cookie_instrument and save_javascript to true to store javascript cookies table, regular (profile) cookies table and all javascript snippets loaded and executed by website respectively.

I repeated each HPW crawl 5 times per persona and 2 times per baseline, to account for infrequent, but unavoidable network errors. In later sections, I label a persona visiting a left- or right- HPW , using 'L' and 'R' respectively, and append this with the persona's acronym. For example, 'L:Y' is for a *Youth* persona visiting *left* HPWs ; similarly 'R:SM' is for a *Senior-Man* persona visiting *right* HPWs .

Since each browser instance is independent, multiple browser instances can be initiated on a crawling server (depending on its available resources). Note that while performing multiple crawls and resetting cookies, some browsers can share same IPs address. Hence, such sessions may not be completely independent of each other. In these cases, based on the server information they may receive similar cookies. This parallelization allows for multiple, simultaneous crawls of personas and baselines, scaling the auditing to many websites at the same time from one location. This also allows us to capture the same tracking and advertising effort from third parties, before ad-campaign budgets change.

4.5.2 Crawling on USA news websites

These crawls were executed in Nov-Dec 2018. The crawling engine was set up on 10 Amazon Web Services (AWS) instances (each: 1 GB memory, 1 core, 8 GB storage, Ubuntu 14.04) in the USA, for instantiating parallel crawls. The crawling was orchestrated by OpenWPM,

Table 4.3 Crawls for persona building (P:) and visits to HPWs (HPW:). Second column: total count of requests or responses across crawls; Third column: average count per website; Fourth: standard deviation (SD) per website; Fifth: median count per website.

Response Type	Total Count	Mean per website	SD per website	Median per website
P: HTTP Requests	53.2k	190	186	135
P: HTTP Responses	49.6k	177	155	132
P: HTTP Redirects	7k	25	46	9
P: Cookies	36.8k	131	192	67
HPW: HTTP Requests	8.3 M	159	194	98
HPW: HTTP Responses	8.1 M	144	181	93
HPW: HTTP Redirects	2.5 M	65	149	40
HPW: Cookies	3.5 M	60	132	16

loading a Firefox 52.0 browser instance and visit websites using Selenium. Each AWS instance uses a parallel and independent OpenWPM instance loaded with a different persona profile. However, in doing so some instances may have shared IPs. Each crawl gives us ∼500MB SQLlite database across the 10 instances. This database stores information about the HTTP calls, cookies, visit sequence, and other meta information based on the settings described earlier.

I implemented the methodology into the framework illustrated in Figure 4.1 for user tracking on 667 HPWs . Each website is marked as left- or right-leaning (W^L and W^R, respectively) based on the description and self-attestation from the website's 'about' page, or facebook page description. Interestingly, out of 667 websites curated in 2016, 111 websites were not active in Nov-Dec 2018 when the crawls took place. Thus, in my study I use 556 active websites: 164 left- and 392 right-leaning HPWs . I normalize my results to account for the imbalance between the two lists. I visit HPWs using baseline (null) and persona profiles. I store the cookies served to each persona profile or new user and analyse their distributions, type of first or third party sending them, popularity of the HPW involved, etc. A summary of my dataset regarding HTTP calls and cookies set during persona building and final crawls for all personas and HPWs is given in Table 4.3.

4.5.3 Wikipedia, the Free Encyclopedia

Following are details of the dataset of Wikipedia articles for the 650 MPs elected in the 2017 general election that I collected as a part of this study.

Page Views To understand readers' engagement, I collect the daily page views data on all articles, using Wikimedia's [5] page view API.[6] I use the earliest possible day that can be set (i.e. 01 July 2015) in the Wiki API query, and obtain daily page views per MP page until 30 June 2019. In total, I observe over 160M views for 650 pages during this period.

Page Edits I crawl the history of page edits for all 650 MPs from 01 June 2002 to 28 Aug 2019. I store in total 231k edits. These edits are made by 43k unique editors. Across all edits I see that 55k edits are made using public IP addresses, which is shown as the username for anonymous editors.

Page Content To understand the content of the pages, I also collect page text as HTML dumps as on 18 July 2019. From these dumps, I extract the paragraph text, section titles, the citations list and other metadata for each page.

MPs Information I obtain additional profile information of MPs using Wikidata.org, a free Wikimedia foundation knowledge base. Wikidata provides information about MPs' gender, party, year of page creation and position held in the Parliament. I identify the role of each MP in politics, which is Wikidata's *Position held (P39)* property. For example, for the current prime minister *Boris Johnson* his positions held include: *Mayor of London, Secretary of State, Member of Parliament, Prime Minister and Leader of the Conservative Party*.

Additional Data I obtain additional baseline data from Wikipedia and Twitter. I collect MPs' interactions on Twitter, such as number and popularity of mentions across 2 months from [ASW19]. For Wikipedia, I crawl page views for *Sportspeople* (footballers) and *actors*. These categories are popular biographies of living people in the UK[7]. I randomly sample 1000 pages from a Wikidata page list of these two categories.

[5] The Wikimedia foundation host projects and websites such as Wikipedia.
[6] https://tools.wmflabs.org/pageviews/, Accessed 26 Feb 2020
[7] www.wikidata.org/wiki/Wikidata:Living_people/uk, Accessed 15 Feb 2020

I summarise the distribution of the dataset in Figure 4.3 and use this to present findings in Section 4.8.

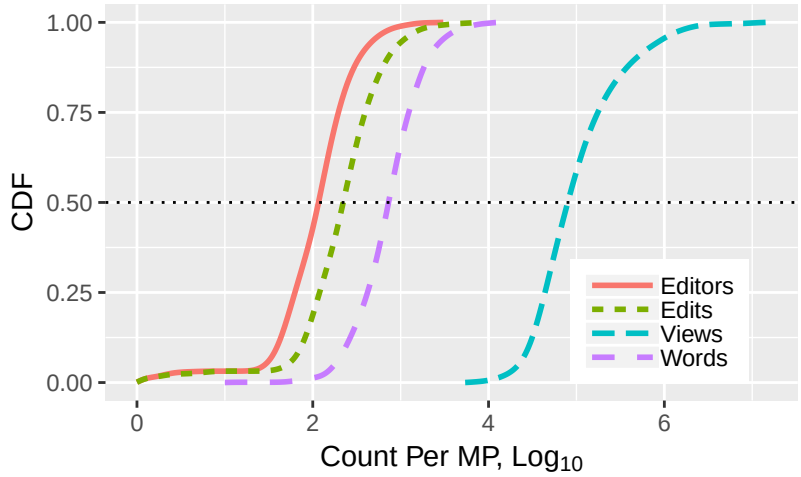


Fig. 4.3 CDF (Cumulative Distribution Function) of number of views, edits and editors per MP.

4.6 User Tracking in HPWs

My modus operandi is to use the previous methodology with above deployed framework to visit various HPW , and examine differences in the tracking performed by the first and third parties included in each. In this section, I focus on null profiles to measure basic tracking performed on new (null) users. I conduct an investigation on the differences in tracking, by comparing left and right HPWs (Sec. 4.6.1), by looking into the top trackers on the Web and whether they are over/under represented in left and right HPWs (Sec. 4.6.2), and finally, by studying popularity of HPWs and checking if this associates with differences in tracking (Sec. 4.6.3).

4.6.1 Who Facilitates More Tracking: Left or Right?

The volume of cookies in a website is a good proxy for measuring how much the website tracks a user directly, and how much it enables third parties in its page to track this user.

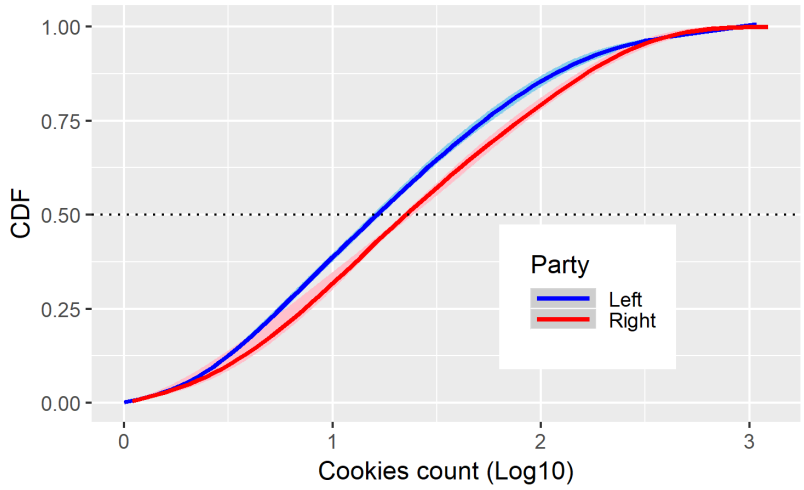


Fig. 4.4 CDF of total number of cookies stored per website, when a baseline user visits left and right-leaning HPWs .

To understand how much of this tracking is happening on HPWs , in Figure 4.4, I plot the Cumulative Distribution Function (CDF) of number of cookies set on a baseline user's profile when visiting W^L and W^R. On average, W^R set 9 more cookies than W^L on baseline users.

Next, I study the type of cookies set on baseline users, using a list of `Disconnect.me` [Dis13]. In Figure 4.5, I breakdown the cookies into six types: *first-party, advertising, analytics, content, social, and other*. Third party cookies from advertising entities significantly outnumber all other types. Furthermore, W^R place more cookies of all types, and especially for advertising, in comparison to W^L.

4.6.2 Is this Tracking More than the General Web?

Earlier, I showed that HPWs enable tracking of users by third parties, and that W^R do so in higher intensity than W^L. But who are the top trackers in these partisan websites? And how different is the tracking they do, when a baseline user visits such HPWs ? I extract the top trackers on the Web using a live list maintained by `whotracks.me` [wbCG19] on 25/09/19, and compare this list against the trackers detected in the two groups of websites, when they are visited by a baseline user. I find that 72% of third parties match between the list and the ones on HPWs . However, there are also differences among the HPW tracking ecosystem and

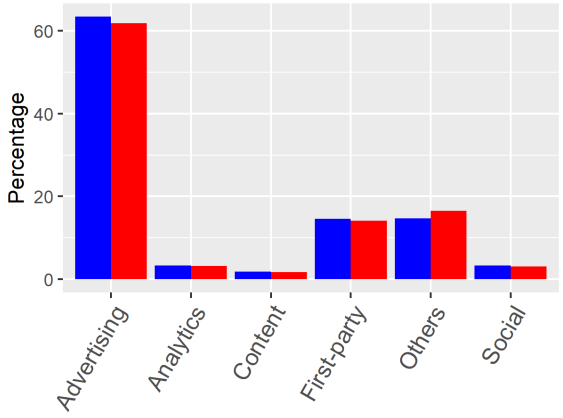


Fig. 4.5 Average number of cookies per website, per type of domain sending them (using the `Disconnect.me` categories).

the Web with respect to intensity in tracking from each party. Figure 4.6 shows a histogram of percentage of websites that specific trackers appear in, and drop cookies on their users. Many of these top trackers are over-represented in the HPWs examined, in comparison to the Web. For example, *Doubleclick, Scorecardresearch, Quantserve, and Adnxs*, appear in at least twice as many websites (proportionally) than in the general Web. Furthermore, many of them appear in many more websites on the right (W^R) than the left (W^L), demonstrating an interest in tracking users visiting right-leaning HPWs .

4.6.3 Is Tracking Associated with Site Popularity?

I observed differences between W^R and W^L in number of cookies they drop on users, type of third parties involved, and top trackers in the Web that also appear in these websites. Next, I investigate how the tracking intensity of HPWs associates with the popularity of each HPW . In Figure 4.7, I present the cookie counts of HPWs dropped on baseline users, against the respective rank range of each website based on Alexa. Again, I find that regardless of the rank range, W^R serve more cookies than W^L. Interestingly, the top ranked W^R (i.e., 1-10K) demonstrate the highest median count, with that count decreasing for lower ranks.

In the next section, I compute various measures to provide statistical evidence of differences in personas and their demographics vs. intensity of tracking facilitated by Hyper-partisan websites .

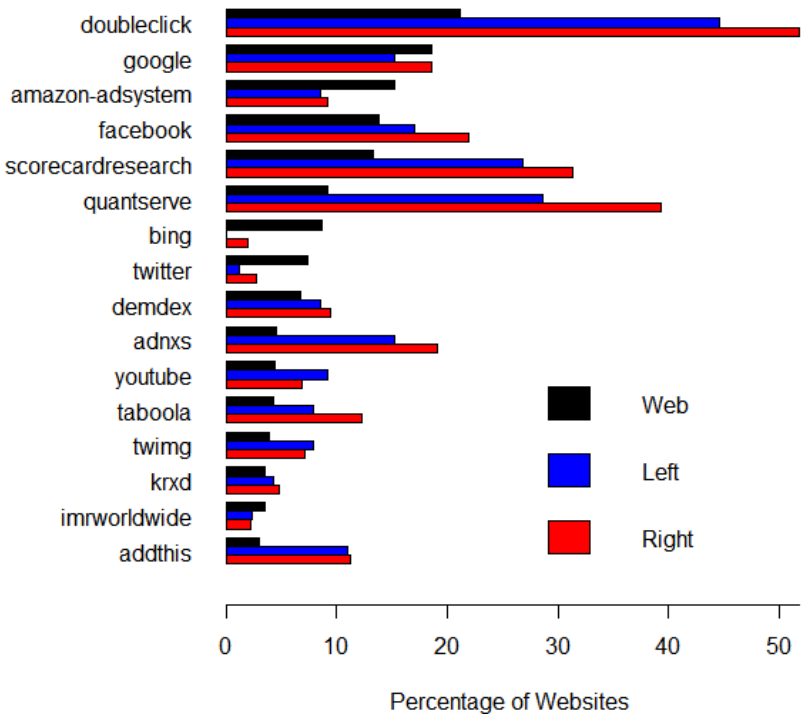


Fig. 4.6 Percentage of websites with cookies set (x-axis) from top 16 tracking domains on the Web (y-axis), for left and right-leaning HPWs , as well as in the Web, for a baseline user. Domain ranking based on live list of `whotracks.me`.

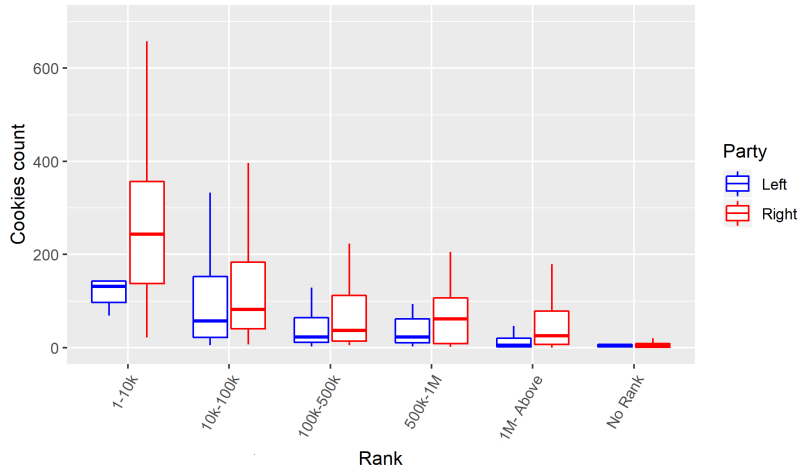


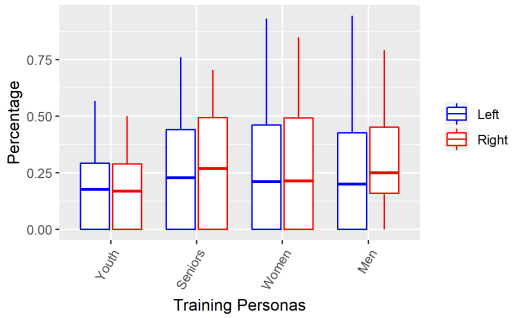
Fig. 4.7 Cookies count for baseline users visiting W^R and W^L, vs. the `alexa.com` global ranking of these websites captured on 25/02/19. Top rank W^R have highest median of cookie count and the median decreases as the rank increases. Websites with no ranking have very few cookies.

4.7 Case Study 1: Preferential Tracking of Personas on Hyper-partisan Websites

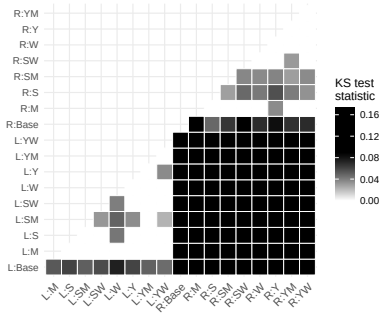
In first case study, I test my framework on self-declared left or right leaning websites [BJBS18]. I hypothesize that the trackers being deployed on these HPWs would converge onto preferentially tracking personas conducive to the prime audiences of these websites. To test this hypothesis, I use Non-Negative Matrix Factorization (NMF) [LS99] to cluster the websites, as well as the targeted demographics into the most viable clusters. The members of these clusters exhibit more similar behaviours to each other, compared to members outside these clusters.

NMF decomposes any non-square matrix A into two components such that $A \approx B \times C$, where B is called the factor *basis* matrix with the dimensions $p \times k$, and C is called the *coefficient* matrix with dimensions $k \times c$, for any choice of k clusters. For a given choice of k, the NMF algorithm tries to solve the optimization problem:

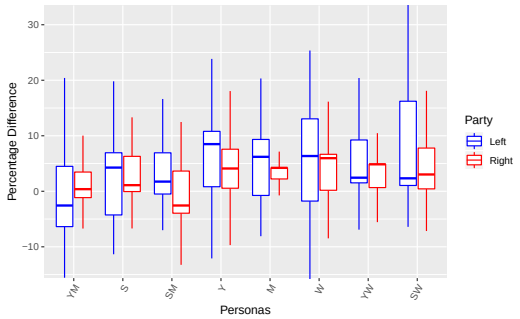
$$min||A - BC||_F^2, \; such\; that\; B \geq 0\; and\; C \geq 0 \tag{4.1}$$



(a) Third parties & persona building third party overlap



(b) KS-statistic for comparison of third parties



(c) Third party difference of baseline & personas

Fig. 4.8 (a) Overlap between third parties dropping cookies during persona building, and when the same personas visit HPWs . (b) Heat-map of KS statistic test for all pairwise comparisons between distributions of numbers of unique third parties serving cookies. All cells with $p \geq 0.01$ are whited out; only cells with $p < 0.01$ are coloured. (c) Percentage difference between third parties serving cookies to baseline and loaded personas. The x-axis is sorted on medians of all personas.

where F represents the *Frobenius* norm. To setup this optimization problem, I first a compose the matrix A, such that it reflects the cookie profile observed by any given persona. I create a $p \times c$ dimensional matrix, which I call A, where p represents the rows corresponding to all the personas I curate, and c represents the columns which correspond to the different third party. I normalize each row, such that A_{ij} now represents the percentage of cookies injected for a user of the P_i persona by the j^{th} domain.

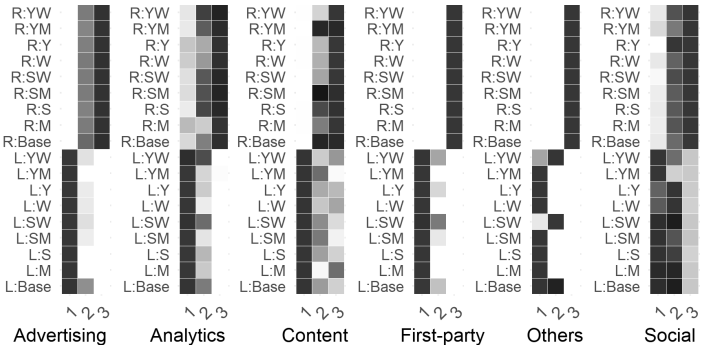
As described in Section 4.6.1, I have distinct set of cookies for each persona viz. *first-party, advertising, analytics, content, social and others*. Hence, I create 6 distinct matrices corresponding to the 6 different types of cookies. I further factorize each of these

Fig. 4.9 Variation in cookies count for foxnews.com (top ranked right leaning) and msnbc.com (top ranked left leaning) with various personas. In both websites, the cookie count is lowest for the baseline persona. X-axis is sorted by increasing count of cookies on both sides.

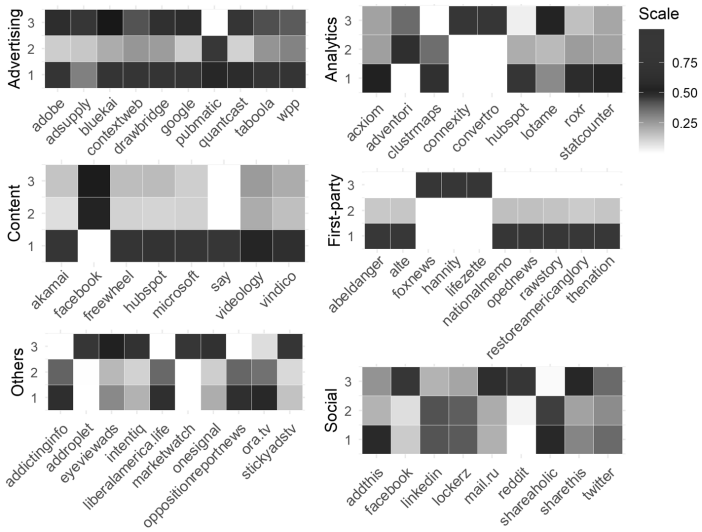
'persona cookie profile' matrices A^t, into corresponding basis B^t and coefficient C^t matrices where $t \in (1,6)$ using NMF.

According to [BTGM04, HMSG08], clusters should be chosen to have high cophenetic correlation and small residuals. After experimenting with various values of $k(\in [2,10])$, I find that the best clustering is obtained when $k = 3$, giving us the maximum cophentic correlation (0.982) and minimum residual value (9).

The results of this clustering have two implications. First, the well-defined clusters for $k = 3$ validate my hypothesis that the trackers deployed on HPWs preferentially converge onto specific partisan personas. Second, it allows us to further pinpoint how left- and right-leaning personas are associated with different kinds of third parties, ranging from advertising to analytics. The value of $k = 3$ also matches the intuition that two of the three clusters could be assigned to the left and right partisan positions, respectively. The third cluster can be interpreted as the middle ground (if any) between party lines. This can be seen in Figure 4.10. Figure 4.10a (first party) shows that the profiles clearly fall into either the Left or Right cluster. Figure 4.10b (first party) shows only the top 10 websites for clarity (rather than all HPWs). Here again, I can see websites fall into either the left or right category. After I computed the NMF on the different matrices A^t, I get the various B^t and C^t factors in



(a) B Factor



(b) C Factor

Fig. 4.10 B^t and C^t factors from NMF clustering of A^t for six categories and 3 clusters (i.e., $k = 3$). I show the scale of all the figures in top-right corner, which is normalized from 0 to 1.

Figure 4.10, for advertising, analytics, etc. Figure 4.10 a shows that 'first party' and 'other' domains essentially take clearly the extreme far clusters, as they have in cookies set from W^L and W^R (Figure 4.10b). Personas exposed to W^L are easier to cluster for advertising, analytics and content, whereas social is less clear. Furthermore, from the C^t factors in Figure 4.10b, I see that some first party domains are clearly clustered in the left (e.g., *The Nation*) and others as right (e.g., *Fox News*). Also, for analytics and advertising, many domains are heavy on the right-wing cluster. For content, many domains including *google, microsoft* and *adobe*

appear to be clustered on the left-wing cluster, apart from *oberon*[8]. Interestingly, for social, *Facebook* and *reddit* are clustered in the right-wing cluster.

4.8 Case Study 2: Quality and Dynamics of Wikipedia Pages about UK Politicians

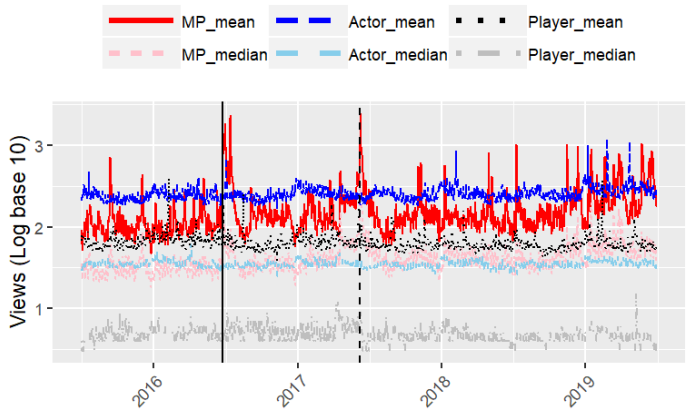
Wikipedia is one such major source of information that provide large variety of content online. Readers go to Wikipedia to get reliable information about different subjects, one of the most popular being living people, and especially politicians. While a lot is known about the general usage and information consumption on Wikipedia, less is known about the life-cycle and quality of Wikipedia articles in the context of politics. The aim of this case study is to quantify and qualify content production and consumption for articles about politicians, with a specific focus on UK Members of Parliament (MPs).

Given the visibility of Wikipedia, and the importance of the online encyclopedia in forming public opinion, the integrity and completeness of its content is crucial, especially during "times of shock"such as elections or referenda [ZWB+19, KSPC07]. To ensure information quality, Wikipedia editors operate in compliance with the core content policies of Neutrality and Verifiability[9]. However, the scale of information and the free-to-edit Wikipedia policy sometimes limit the capability of communities to maintain the quality and neutrality of Wikipedia pages, and little is known about the life-cycle and quality of Wikipedia articles in the context of politics. I next present findings and results based on the collected dataset.

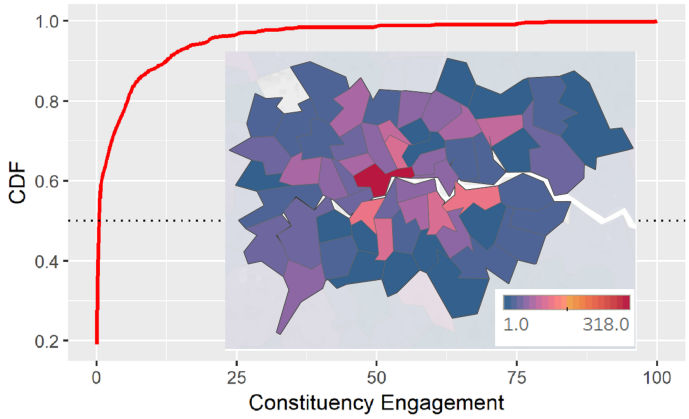
Spatio-temporal patterns of Pages

I begin by investigating when the MPs' Wiki pages are created and edited and what events or actions have impact on page views. In summary, I find that most MP pages are created soon after they are first elected; edits happen after significant changes such as elections or

[8] oberon-media.com is a multi-platform firm that gives gaming solutions
[9] https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies



(a) Page Views



(b) Constituency Edits

Fig. 4.11 Spatio-temporal patterns. (a) Page views of MPs pages and baselines (footballers and actors in the UK). (b) CDF of constituency engagement. Inset: All edits from constituencies of Greater London.

scandals; and most views happen just before or after important events such as referenda, elections and scandals.

Using edit history for each article, I obtain the article creation date (i.e. the day when the first edit was made). The edit history data covers four UK general elections (2005, 2010, 2015 and 2017). I see from that for the majority of pages, the creation date is close to a general election (60% of page creation falls within these four years). In addition, 25% of the articles were created between 2002 and 2004 (both included), coinciding with the birth and subsequent rapid growth of Wikipedia. I focus on the dynamics of viewing behaviour. MP pages obtain a large number of views, with an average of 247k views per MP during

the period I consider. I show the average view count of each day (mean) and median in Figure 4.11a (Note: Y axis is log scale). The mean views count is high on days of two major events: the *UK EU membership referendum* (also known as the 'Brexit' referendum) of 24 June 2016 and the *UK general election* of 09 June 2017. Apart from events where a majority of MP pages get attention, there are events which are specific to individual (non-popular) MPs when they are in the limelight. Examples include anticipated and unanticipated events, such as ministers' resignations, speeches, interviews etc. [ASW19].

Another core dimension of engagement with Wikipedia is the editing process carried out by largely anonymous volunteers. Among all edits, 36% happen during the last three election years captured in the dataset (2010, 2015 and 2017). These patterns are similar to the views pattern in Figure 4.11a (Pearson's correlation: 62%, $p < 0.001$), and hence not shown. Out of all the edits, around 55k (22%) edits are by public IPs which are recorded in the page revision history for not logged-in (anonymous) users. To understand the spatial distribution of editors, I map each public IP to a possible physical location (postcode), using the service provided by *db-ip.com*, a geo-location database. I observe that 84% of public IP edits are from the UK. The remainder are from countries such as the United States–1768 edits, Ireland–442 edits, Australia–364 edits, Canada–297 edits, etc.

I then map each postcode to a local constituency using the list provided in [ASW19]. I plot edit location at constituency level in Figure 4.11b (inset), and observe that most public IP edits come from the area of Greater London, and more specifically from the Westminister borough (318 edits) where the Parliament sits. This is indicative of MPs' staff managing edits, similar to the findings of previous studies, which could provide a source of bias in the articles [GM18]. The highest number of public IP edits coming from an MP's local constituency is for MP *Amber Rudd MP*, former home secretary, who has 47 out of 232 edits from her own constituency (Hastings and Rye).

To further quantify the extent to which edits come from an MP's constituency, I calculate the *Constituency Engagement Factor (CE)* of each MP m as the proportion of the edits to their articles which are localised to their constituency. If m has N_m edits of their page and

$N_{m,c}$ edits from their constituency, I write:

$$CE_{m,c} = \frac{N_{m,c}}{N_m} \tag{4.2}$$

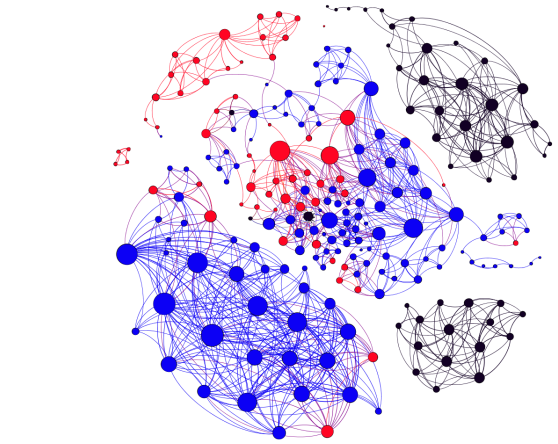
I plot the distribution of the metric in Equation 4.2 in Figure 4.11b. I see that for 31% of MPs there is at least one edit from their own constituency, but *CE* is in general low, and only 6 MPs have more than half of their edits from their own constituency.

Polarisation

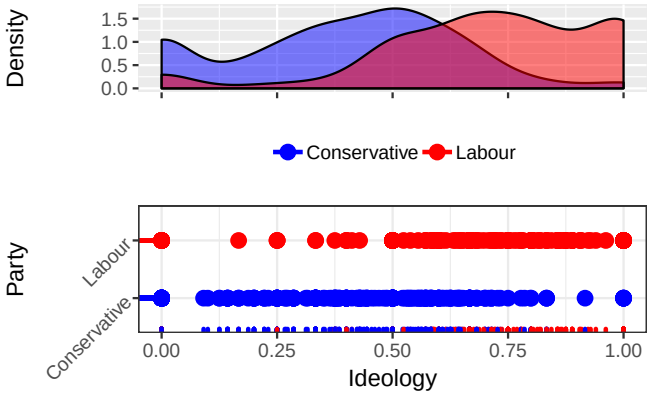
I continue this case study by looking at information quality through the lenses of potential ideological and societal biases in Wikipedia articles about UK MPs. I ask – Do editors have an ideological bias, e.g., focusing on editing pages of MPs from a specific party? Do they have an ideological slant, and is their coverage sufficient? I find that there is a specialisation of editors, with some focusing mainly on Conservative party MPs, others on Labour MPs, and so on.

To understand the extent to which editors tend to polarize around a specific ideology, I start by tracing, for each of the 42k editors, the party of pages which they mostly edit. To this aim, I compute the number of edits to MPs from a given party, similar to [Bar15, ASW19], and associate each editor to the party to which they contribute the maximum edits. I do this for editors editing at least three different MPs, in order to exclude cases in which an editor is only interested in one MP or two MPs from different parties. This filtering step leaves the processed data with 4.2k (7%) of editors who are collectively responsible for 67% of edits. Note that the vast majority of editors (nearly 95%), even when they edit more than one MP pages edit pages from only one party. Such editors are considered to be supporters of that party. A minority of users (5%) edit MP pages from more than one party but in most cases this interest is unequal. The editor is then considered to be a supporter of that party.

To understand communities forming this polarisation I perform network analysis – I define editors as nodes and induce weighted edges between each pair of editors by computing the *Jaccard Coefficient* or similarity of the sets of pages edited by them. Thus, if two editors edit exactly the same set of MP pages, the weight will have its highest value of 1, and if there



(a) Editors Network



(b) Ideology scores

Fig. 4.12 Measure of Polarisation (Red: Labour, Blue: Conservative, Black: Others). (a) Network graph based on editors as nodes, with edges connecting editors who have edited the same MPs' pages. (b) Ideology scores and density based on citations domains.

is no overlap, the value will be 0 (considered as no edge). I then use the Louvain method to identify communities of editors who have more connections within each community, but not many connections across communities. I find that this graph cleaves into 8 tightly knit communities, with a moderate modularity score of 0.229 [NG04], indicative of polarisation or specialisation by party among editors.

To visualise this better, I focus on the most active editors, and remove nodes with a degree of less than 5. To remove clutter, I do not show edges with weights less than 0.5. Figure 4.12a depicts this graph by colouring the nodes (editors) based on their party (Blue

for Conservative editors, red for Labour and black for others), and visually confirms the polarisation detected above by showing how the graph of editors divides along party lines,.

I next focus on the *content* rather than the authors. One may expect that if authors exhibit polarisation, the sources they draw from to write the content, i.e., the MP pages, may also be polarised. To quantify this, I use URLs which are embedded in the *References* section as citations. I find that there are 19k citations on MP pages. The average citation count is 29 per page and the median is 19. By checking the top citation URLs, I see that majority of them are news domains. This is consistent with the English Wikipedia articles study by [PRCW20] which shows that top domains cited are Google.com, DOI.org, Nytimes.com. NIH.gov, BBC.co.uk, TheGuardian.com and so forth. Similar to [PRCW20] I also extract the base domain from each citation URL and obtain 2212 unique domains. Additionally I find long URLs for 1.4k web archive *(web.archive.org and archive.is)* short URLs. The top 5 domains which I get are BBC (15%), theguardian (11%), telegraph (6%), parliament.uk (6%), independent (5%). The top 10 domains cover 52% of the citations list.

To check the use and representation of news domain sources on MP pages I compute an ideology score for each MP page. To do this I first check and label all possible news domains with their political leaning. I use scores from *Mediabiasfactcheck* [Med20] and label the top 50 news domains with their political leaning (left, center or right). I also add ideology scores for news domains using sub-strings such as conservative (right-leaning) or labour (left-leaning) in the domain name. With this approach I find and label 67% (13k) of the domains. Some examples of top domains by citations count are Left–Theguardian (2069), Independent (969), Newstatesman (262); Center–BBC (2766), Parliament.uk (1151), UKwhoswho (233); and Right–Telegraph (1207), thegazette (302), thetimes (251).

After labeling, I compute an ideology score for each page, as the fraction of identified and *Mediabiasfactcheck*–labelled news domains on that page that are left-leaning (i.e., ideology score = (number of left leaning news domains in citations)/(number of left+right leaning citations)). Note that a score of 0 indicates usage of only right-leaning sources and a score of 1 indicates only left-leaning sources; a score of 0.5 indicates a perfect balance. Figure 4.12b shows the ideology scores of Conservative and Labour pages as scatter and density plots.

Interestingly, two peaks are seen (median Labour: 0.7 and median Conservative: 0.4) which indicate a slight polarisation, with a slightly more polarised (i.e., farther away from 0.5) score for Labour MP pages. Also, 53 (17%) Conservative and 57 (22%) Labour pages have extreme polarised scores of 0 and 1 respectively. The KS statistics test (two sample) also confirms that there is a significant difference ($p < 0.001$) with a distance value $D = 0.54$ in the two parties' ideology scores.

4.9 Discussion

Understanding the dissemination of misinformation over online platforms and fringe news sources (e.g., Breitbart or Infowars) has become a crucial topic of study because of its considerable impact on online culture. These sources are often found responsible for spreading news and opinion pieces with extreme views along partisan lines. This strategy keeps users engaged with the websites for prolonged periods of time. A major incentive of such Hyper-partisan websites (HPWs) to follow this strategy is monetisation by delivering targeted and personalized ad-content to profiled users. In fact, these websites do not need to provide any credible content or news, thus, reducing costs for journalists and other expenses, otherwise incurred on legitimate news websites. However, until now, it remained unclear how such HPWs perform targeted advertising and whether they *differentiate* their online user tracking based on demographics and partisan leaning of their audience.

To shed light to this problem, in this chapter, I present a methodology that creates artificial users (personas) with certain demographic features. I find that a stateful browsing of a small number of 10 websites is sufficient to build a persona (cookie state) of hundreds of unique third parties. I store browser states of personas and use them to emulate diverse types of browsing patterns along partisan lines, and in the process, record how these websites track them. The data acquired from such persona-based crawls are an asset for investigating user experience, bias in tracking, and various other privacy metrics. By visiting 556 HPWs , I collected and analysed a dataset of $\sim$19 million HTTP (requests, responses and redirects) calls and $\sim$3.5 million cookies. In general, I find a higher amount of tracking in the right-leaning

HPWs. Note that I did not investigate the content of cookies as part of this study. This can be done in future studies using the data of this research. Also, in this work I only visit the homepage (main page) of every website. A deeper examination of sub-pages using much larger data crawls could help in understanding topical tracking within a website.

Among the right-leaning HPWs , the top ranked (by Alexa) websites are consistently loaded with more trackers, in comparison to less popular HPWs . By performing a co-clustering (first case study) on the personas using the NMF model, I observe that I can group personas and third parties in a low dimensional space using the types of cookie domains as features. I obtain three optimal clusters which can be labelled as (mild) left, (mild) right and neutral.

In second case study, I discussed important dynamics of attention and content polarisation in the context of Wikipedia articles focusing on UK Members of Parliament. I find evidence of specialisation in contribution patterns of the editors of MP pages. With this analysis, I contribute to the broader field of online political communication studies, and shed light on behaviour of contributors to the largest online encyclopedia. Similar to previous work, I find that MP page creations, views and edits are strongly aligned with media coverage and election periods. Furthermore, I find that only a small fraction of edits come from within the constituency of the MP, whereas the majority of anonymous edits come from Central London where the political centre of the UK lies. Collectively, these findings suggest that attention peaks are localized both in time and space, thus introducing potential vulnerabilities to the integrity of the content. Researchers working on monitoring and detecting coordinated disinformation attacks in political communications might benefit from these findings and further investigate how these peaks of attention might affect temporarily the quality of content on Wikipedia MP pages.

For the benefit of reproducibility, I reported an in-depth process for the creation and deployment of said personas. This approach is easily scalable across different demographic features and browsing patterns. My framework can also be repurposed for auditing other ecosystems such as e-commerce, web search engines, public social media pages, and other recommendation services. My framework can be useful for enforcing new regulations under data protection (e.g., GDPR [Eur18]) which have been introduced to give back control of

personal data to their owners, and increase overall Web transparency. As GDPR enforcement is in its initial phase, most of the control on user's data and tracking still remains with the websites, allowing many HPWs to take liberties on how to comply with regulations. For example, in this study, I found cases of right-leaning HPWs (e.g., spectator.org) and left-leaning HPWs (e.g., newshounds.us) setting more than 1000 cookies per user.

CHAPTER 5

INTERACTIONS BETWEEN CITIZENS AND DIGITAL PLATFORMS

Be the change that you wish to see in the world.

Mahatama Gandhi

5.1 Introduction

Soon after its independence, India was divided internally into states, based mainly on the different languages spoken in each region, thus creating a distinct regional identity among its citizens in addition to the idea of nationhood [Guh17]. There have often been conflicts and disagreements across state borders, where separate regional identities have been asserted over national unity [Gup70]. Within this historical and political context, I wish to understand the extent to which information is shared across different languages.

In this Chapter, I take as a case study the 2019 Indian General Election, considered to be the largest ever exercise in representative democracy, with over 67% of the 900 million strong electorate participating in the polls. Similar to recent elections in other countries [MM12, ASW19], it is also undeniable that social media played an important role in the Indian General Elections, helping mobilise voters and spreading information about the different major parties [Rao19, Pat19a]. Given the rich linguistic diversity in India, I wish to understand the manner in which such a large scale *national* effort plays out on social media despite the differences in languages among the electorate. I focus my efforts on

understanding whether and to what extent information on social media crosses regional and linguistic divides. I characterise this as interactions between citizens and platforms in newly emerged multilingual platform–Sharechat and WhatsApp [1] using dataset which are shared by collaborators.

My key hypothesis is that these language barriers may be overcome via different kind of contents, which is less dependent on linguistic understanding than text or audio. To explore this, I formulate the following two research questions:

RQ-1 Can image content transcend language silos? If so, how often does this happen, and amongst which languages?

RQ-2 What kinds of images are most successful in transcending language silos, and do their semantic meanings mutate?

RQ-3 What are the emerging challenges which platforms need to consider for tackling misbehaviour and maintaining political neutrality?

To answer RQ-1 and RQ2, I exploit the ShareChat data and propose a methodology to cluster perceptually similar images, across languages, and understand how they change per community. Exploring RQ-1, I find that a number of these images are contain associated text [ZCB+18]. I use the extracted language features within images and further annotate multi-lingual images with translations and semantic tags. Using this data, I investigate the characteristics of images that have a wider cross-lingual adoption. I find that a majority of the images have embedded text in them, which is recognised by OCR. This presence of text strongly affects the sharing of images across languages. I find, for example, that sharing is easier when the image text is in languages such as Hindi and English, which are lingua franca widely understood across large portions of India. I also find that the text of the image is translated when shared across different language communities, and that images spread more easily across linguistically related languages or in languages spoken in adjacent regions. For RQ2, I present a case study 1 in Section 5.5. In this case, I observe that sometimes message

[1] Interactions on WhatsApp platform are discussed as a case study in Section 5.6

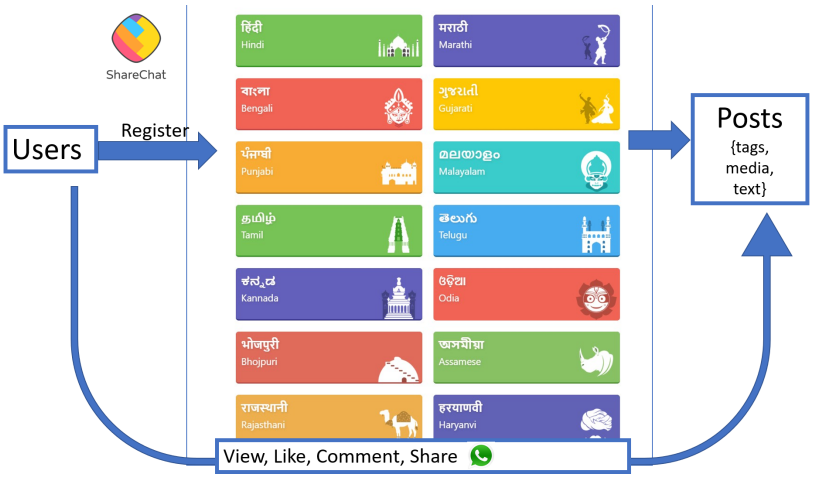


Fig. 5.1 ShareChat homepage and user interactions.

change during translation, *e.g.,* political memes are altered to make them more consumable in a specific language, or the meaning is altered on purpose in the shared text. These results have key implications for understanding political communications in multi-lingual environments.

For RQ-3, I focus on a peer-to-peer messaging platform– WhatsApp . With massive popularity of WhatsApp , *unwanted junk messages* have also become a challenge. However, unlike spam email, where platforms can read content, WhatsApp follows an end-to-end encryption model where the message content is not accessible. Although this offers stronger guarantees on privacy, it makes moderation and spam detection difficult. Although WhatsApp has made progress [Jon17] in detecting users who send unsolicited messages to individuals, there is no solution for spammers who send to *public WhatsApp groups* [GT18]. In Section 5.6, I present a case studies on tackling such spamming misbehaviour in political chat groups.

5.2 Background and Data Collection

5.2.1 Multi-lingual Social Media: Sharechat

Overview of Sharechat- Made in India

Figure 5.1 presents the ShareChat homepage, in which users must first select the language community into which they post. Note that there is *no* support for English language during the registration process,[2] thereby creating a unique social media platform for regional languages. Thus, by mirroring the language-based divisions of the political map of India, ShareChat offers a fascinating environment to study Indian identity along national and regional lines.

The dataset is shared through collaboration with researchers from MIT, USA. The collection of this dataset started with the set of manually curated topics that Sharechat publishes everyday on their website, for each language.[3] These topics are separated across various categories, such as entertainment, politics, religion, news, with each topic containing a set of popular hashtags. In total, 1,313 hashtags were used, of which 480 were related to Politics. Thus, I believe Politics represents a significant portion of activity on Sharechat during the period of study. Note, that the period I consider coincides with high profile events of national importance, such as the Indian National Elections and an escalation of the India-Pakistan conflict.

In functionality, Sharechat is similar to services like Instagram, with users sharing mostly multimedia content. However, it has one major difference that aids my research design: Unlike other global platforms, different languages are separated into communities of their own, which creates a silo effect and a strong sense of commonality. Despite this regional appeal, it has accumulated tens of millions of users in just 4 years of existence, with most being first time Internet users [Lev19].

Data Collection Methodology

To explore this, I use the collected data consisting of over 1.2 million posts across 14 languages during and before the election campaigning period, from **ShareChat**[4]. This is a media and text-sharing platform designed for Indian users, with over 50 million monthly active users[5]. All the posts from a curated list of political hashtags are fetched using the ShareChat API, for the entire duration of the crawl. Each post is serialised as a JSON object

[2]https://techcrunch.com/2019/08/15/sharechat-twitter-seriesd/

[3]https://sharechat.com/explore

[4]An anonymised version of the dataset is available for non-commercial research usage from https://tiny.cc/share-chat.

[5]https://we.sharechat.com/

containing the complete metadata about the post, including the number of likes, comments, shares on WhatsApp,[6] tags, post identifier, Optical Character Recognition (OCR) text from the image and date of creation as shared by the API of the platform.

I further download all the images, video and audio content from the posts. Overall, the dataset consists of 1.2 million posts from 321k unique users. These posts consist of a diverse set of media types (Gifs, images, videos), out of which almost half (641k) are images. Since images are the most dominant medium of communication, in the following sections, I only consider posts containing images. Images also receive significantly more engagement on the platform, with a median of **377** views per image as opposed to **172** for non-images.

Around 15% of images have no OCR text (i.e. no textual content in the image) in them. I also identify the language from the OCR text and hashtags using Google's Compact Language Detector [Oom18]. I manually checked the accuracy of OCR and language detection on a few hundred posts in multiple languages and report the performance in Figure 5.2. This is based on manual inspection of images with the corresponding OCR generated text.

In the rest of the paper, the term *language* refers to the profile language of the user who authored a post. When referencing languages identified from post processing, I will precede it with the method used to identify the language, i.e., *OCR language* or *Tag language*. I believe that my large scale multi-lingual, multi-modal dataset would be valuable for research in Computational Social Science, and could also benefit researchers from other fields including Natural Language Processing, and Computer Vision.

Data Processing Methodology

I next process the data to (*i*) cluster similar images together; and (*ii*) translate all text into English.

Part 1: Image Clustering Since in this study I am dealing with more than half a million images, in order to make the analysis tractable, I cluster together visually and perceptually similar images. This allows to effectively find "memes" [ZCB+18]. For simplicity, I refer to any images containing accompanying text as memes. To identify clusters of images, I make

[6]Unlike Twitter, ShareChat does not have a retweet button, but allows users to quickly share content to WhatsApp. Hence, the share counts reported here are shares from ShareChat to WhatsApp.

use of a recently open sourced image hashing algorithm by Facebook known as PDQ.[7] The PDQ hashing algorithm is a significant improvement over the commonly used phash [Zau10], and produces a 256 bit hash using a discrete cosine transformation algorithm. PDQ is used by Facebook to detect similar content, and is the state-of-the-art approach for identifying and clustering similar images.

Using PDQ, I cluster images and inspect each cluster for features they correspond to. The clustering takes one parameter d: the distance threshold for the images to be considered similar. This parameter takes values in the range 31–63. Through manual inspection, I find that for $d > 50$, images that are very different from each other tend to get grouped in the same cluster, and for $d < 50$, the choice of d does not appear to make much of a difference, and clusters typically show a high degree of cohesion (i.e., the clusters largely consist of copies of one unique image). Therefore, having tested multiple possible values of the distance parameter, I choose $d = 31$, the default value used in PDQ, as this yields cohesive clusters. Through this process, I obtain over 400k image clusters from 560k individual images. Out of these, 54k clusters have between 2–10 images, and roughly 2000 clusters have over 10 images. The biggest cluster contains 261 images.

Part 2: OCR Language Translation To enable analysis, I next translate all the OCR text contained within images into English using Google Translate. To validate correctness, I ask native language speakers from 8 of the 10 languages I consider, to verify the translations. My validation exercise focuses on 63 clusters (containing 769 images) which are identified as having images with more than 3 different languages, according to OCR. The native speakers first check whether the OCR has correctly identified the text contained in the image, and then check if the machine translation by Google Translate is accurate. In the case of errors in either of these two steps, volunteers provided high quality manual translations to provide an accurate ground truth.

In Figure 5.2 I first present the recorded accuracy of the translation of OCR from the native language to English (orange bar). Overall, this shows that the OCR language detection used by ShareChat performs well (90% to 100%). Figure 5.2 also shows the accuracy of the

[7] github.com/facebook/ThreatExchange/tree/master/hashing/pdq

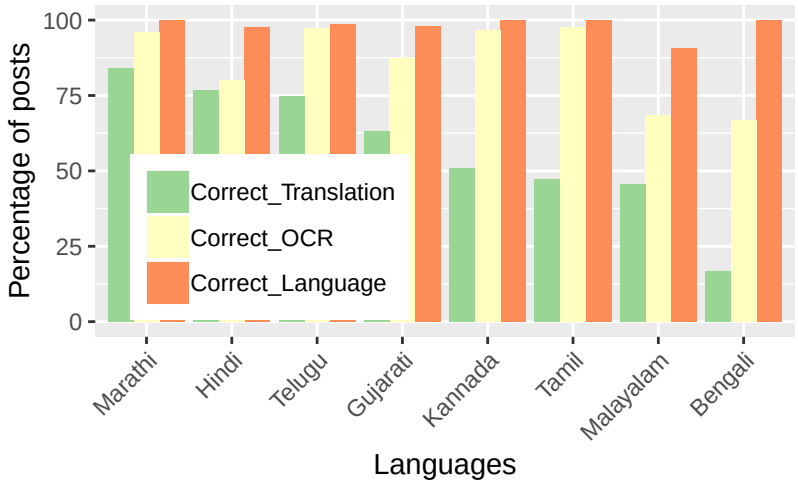


Fig. 5.2 Accuracy of translation and language detection, as evaluated from a sample of 769 images by native speakers: Orange bar shows the percentage of posts for which the language detected by OCR was accurate (close to 100% for all languages). Yellow bar shows the accuracy of the OCR-generated text ($>$ 75% for all languages except Malayalam and Bengali). Green bar shows the accuracy of automatic translation into English.

extracted OCR text (yellow bar). Based on the language, I see mostly positive results (*i.e.,* above 90% accuracy), except for a few, *e.g.,* Bengali which falls under 70%. Finally, the figure shows that automatic translation by Google is not that accurate (green bar). For certain widely spoken languages such as Hindi, translation performs well: over 75% of translations require no modification. However, this does not hold true for many other Indian languages. For example, with Telugu, translation was sensitive to spaces, and performed poorly for long sentences. In contrast, Tamil translations prove to be poorer for short sentences. Furthermore, even minor typographical mistakes in the OCR text negatively impact the results, without the capacity to 'auto correct' as seen with English.

5.2.2 End-to-End Encrypted Messaging App: WhatsApp

Overview of WhatsApp Messenger

WhatsApp is an end-to-end encrypted mobile messaging application. Contacts can be international as well as domestic phone numbers and WhatsApp does not charge for messages. Apart from direct messages, users can join groups where members get messages

and notifications for every post in the group. A user's phone number is their WhatsApp identifier. Partly because of this near zero cost structure, WhatsApp has gained popularity in developing nations and has acquired a large user base.

Due to this scale, anecdotal evidence suggests that malicious actors have started to take advantage of WhatsApp . Particularly, unwanted or unsolicited messages have been on the rise [Bla20, Jon17]. This is dangerous in the Indian context where many WhatsApp users are digital neophytes who may not have experienced digital spam previously. Such spam is particularly prevalent on public WhatsApp groups, where any one can join through an openly available link.

Data collection methodology

Taking inspiration from prior work, which found that public groups that discuss politics are widely used in India [Lok18] and Brazil [NFKN19], I present a case study based on data from *public* WhatsApp groups discussing politics in India. The dataset is shared through collaboration with researchers from MIT, USA. The shared data is collected using an extensive set of 349 manually curated keywords[8] from multiple Indian languages (including English) relating to politicians and political parties in multiple languages. The keywords are searched for WhatsApp group links (`chat.whatsapp.com`) on Facebook, Twitter and Google during November 2018. This yielded 5,051 groups. These are typically created by political parties or party supporters in order to reach an audience which is only available via WhatsApp. Hence, most of these groups have a well defined organisational structure [B$^+$19]. Note, due to end-to-end encryption, WhatsApp is limited in its ability to moderate content in these groups. Instead, moderation is mostly up to group admins, who also have powers to remove users.

Using the toolkit from [GT18] the data is collected from 5,051 selected groups. Using Selenium to programmatically join each group, periodic snapshots of the message is taking and stored by between October 2018 and August 2019. Across the 5,051 groups, there are 2.6 million messages posted by over 172K unique users over a period of 302 days. Also, the data record 437K action events, covering actions taken by users including entering or leaving

[8]List available: https://www.dropbox.com/sh/cm66rha982f2hlj/AADi5QZLIiz0n6iQ9aEVTF9ua?dl=0

Table 5.1 Actions captured within a group.

Action	**Description**	**Action Counts**	**Unique Users**
added	added by a member	61k	37k
added_by_admin	added by a group admin	73k	49k
joined_via_link	joined via an invite link	132k	54k
left	left the group	154k	73k
removed	removed from a group	9k	7.3k
number_changed	changed from one number to another	6k	1.5k

groups and changing phone numbers. Table 5.1 summarizes the actions performed by users within a group. **Ethics note**: The data collection abides by the terms of service of WhatsApp and was approved by the IRB at MIT university. All data was anonymised before analysis, and any personally identifiable information was masked. All phone numbers were one-way hashed, after extracting the country code.

The dataset consists of 2.6M messages out of which only 1.4M are unique. I filter and cluster these messages to have a unique and distinct set of messages as follows.

Message pre-processing: Filtering. Because the content of the message is important in identifying junk, I focus on the four top languages (Hindi, English, Telugu and Tamil) which collectively represent 74% of messages in dataset. I also remove messages containing just URLs (no accompanying text), boilerplate content such as ‘hi’ or ‘good morning’, and, messages consisting solely of emojis, which constitute around 25% of data. I filter such content to avoid characterising low entropy posts – although they are off-topic and widely considered as junk as well [GSAG19], it is relatively easy to implement client-side filters for specific text such as ‘hi’ or ‘good morning’, whereas filtering out the other messages which I classify as junk is a harder problem (studied in Case Study 1–Section 5.5). Filtering these out, the data consists 766K messages.

Message pre-processing: Clustering. I qualitatively observed that many messages are close variants of each other. To group together near-similar variants of the same message, I use MinHash and Locality Sensitive Hashing (LSH) [GIM+99, Mul15].[9]

I find that the best clustering performance is obtained by using 10 min-hashes in 5 bands. I made this choice by taking 100 near-identical messages and 100 distinct messages, and experimenting with a range of parameters to derive the optimal for separating out these two sets.

Overall, this results in 73K clusters, with an average of 10 (median 4) messages in each cluster. As a quality check, I randomly select 100 random clusters from each of the 4 languages and manually verify that 97% are similar (the rest are from a news bot [DUT18], where the end note of the message is identical but the remaining text is disjointed).

Of the 73K clusters, 25.4K have at least 5 messages, and contain 420K messages in total. These 420K messages, which I term "frequently sent messages" (sometimes shortened to "frequent messages") are the object of study for the rest of the paper, except when studying URLs, where I include all the messages containing URLs.

Message pre-processing: Junk annotation The above yields a substantial WhatsApp message dataset. I next describe how I annotate these messages as 'junk' or 'non-junk'.

Part 1 Annotation: Manual annotation. I start by identifying a seed set of users who were manually removed from at least two groups by their admins. I conjecture that these 257 users may be likely to share spam. I then extract the 68K messages (grouped into 1,004 clusters) sent by these users and manually annotate each cluster as 'spam' or 'ham'.[10]

Annotations are performed in two stages. First, two English-speaking annotators work independently to label the 220 English clusters contained within the 1,004 clusters. As guidelines for the annotators, I ask them to identify as spam the following kinds of messages: (*i*) Promotion messages (if non-political promotion as these groups are for political discussions). (*ii*) Adult content related messages. (*iii*) Invitations to register for external services via

[9]Note that other approaches such as fuzzy matching or semantic sentence embeddings could be used here. Given the scale of the dataset, LSH was a reasonable choice. LSH has been used for similar tasks in the past [SKV10].

[10]The term 'Ham' is often used as a synonym for 'non-spam'.

links, often for money. (*iv*) Offers to earn money, win prizes etc. or (*v*) Anything which looks 'suspicious' and not relevant to the group at large. Although this may exclude other more nuanced forms of spam, it offers a powerful lower bound to work from. The inter-rater agreement (Cohen's-Kappa) was high (0.96). The remaining disagreements between annotators were discussed to reach full agreement. Following this initial step, which validated the ability to identify junk messages with high agreement levels, I progress to the second stage. Here, I issue the remaining set of 784 clusters to annotators (native speakers of Hindi, Telugu and Tamil).[11] Each annotator received the subset of messages in their native language, and were given identical guidelines (as described above). In total, the above two-steps result in 63K messages (from 663 clusters) being tagged as spam and 5K messages (341 clusters) being tagged as ham.

To verify that my conception of junk and non-junk is reasonable, I created a panel of 11 independent assessors who looked at a randomly chosen set of 100 messages (split into 50 spam and 50 ham messages). All assessors were Indians, and collectively represented 7 states. The panel were only given the information that these messages are from public WhatsApp groups related to Indian politics and were asked whether they agree with annotators' 'junk' and 'non-junk' labels, based on their own personal understanding of what they would consider as 'junk'. I find that there is a median agreement on 95% of the labels across all the assessors.

Part 2 Annotation :Semi-Automatic Annotation. To broaden my analysis, I construct a dictionary of words that are used at least 5 times in the 63K messages classified as spam above, and manually clean the list to obtain a set of 324 high precision spam words.[12] I then search for the occurrence of these spam words in the entire database of 420K frequently sent messages in Hindi, English, Telugu and Tamil. In each cluster where I find one or more of the spam words, the annotators examine the cluster and validate it as 'spam' or 'ham' using the same approach as above. Finally, I obtain a labelled dataset containing 295K junk (from 3.5K clusters) and 112K non-junk (from 3.2K clusters) messages as summarised in Table 5.2. Some examples of top junk messages are listed in Appendix 7.2.

[11] There were 2 annotators for English, 3 for Hindi and 1 each for Telugu and Tamil.

[12] The threshold 5 was set by manually inspecting the results obtained by selecting various values from 3 to 10 in order to obtain a high precision list of spam words.

Table 5.2 Summary of the annotations set.

Message type	From	Unique messages	Total messages
Junk	Removed users	663	63K
Non-Junk	Removed users	341	5K
Junk	Semi-automation	2.8K	232K
Non-Junk	Semi-automation	2.9K	107K

5.3 Basic Characterisation

I begin by providing a brief statistical overview of activity on ShareChat.

5.3.1 Summary Statistics

Figure 5.3 presents the number of posts in each language. Due the varying population sizes, I see a strong skew towards widely spoken languages. Surprisingly, Hindi is *not* the most popular language though. Instead, Telugu and Malayalam accumulate the majority of posts (16.4% and 15% respectively). Due to this skew, in the later sections I choose to use just the top 10 languages, as languages such as Bhojpuri, Haryanvi and Rajasthani accumulate very few posts.

Next, Figure 5.4 presents the empirical distributions of likes, shares and views across all posts in the dataset. Naturally, I see that views dominate with a median of 340 per post. As seen in other social media platforms, I observe a significant skew with the top 10% of posts accumulating 76% (2.47 Billion) of all views [Sas12, CS12, TESU15]. Similar patterns are seen within sharing and liking patterns. Liking is fractionally more popular than sharing, although I note that ShareChat does not have a retweet-like share option to share content on the platform. Instead, sharing is a means of *re-posting* the content from ShareChat onto WhatsApp.

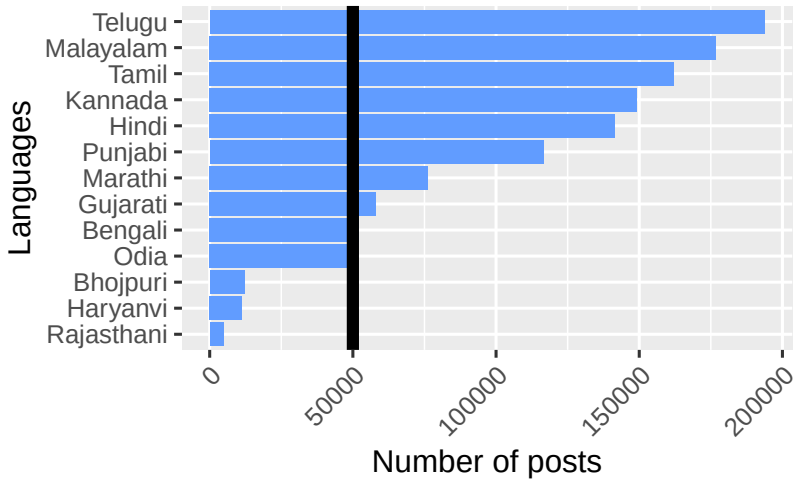


Fig. 5.3 Posts count per language.

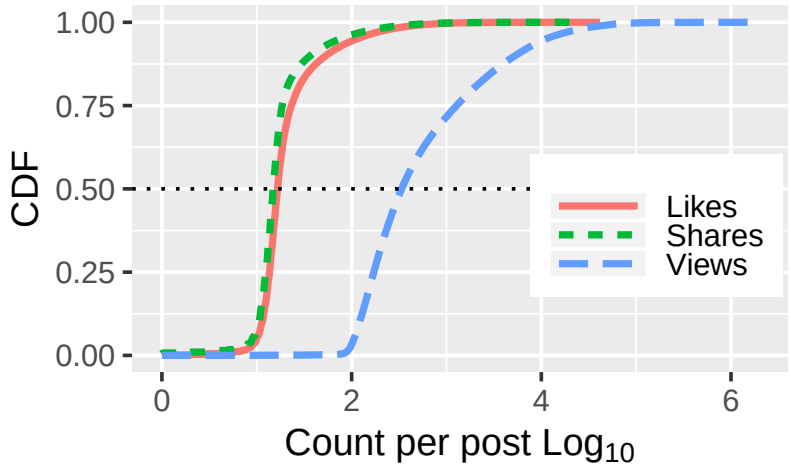


Fig. 5.4 Views, shares and likes across all languages.

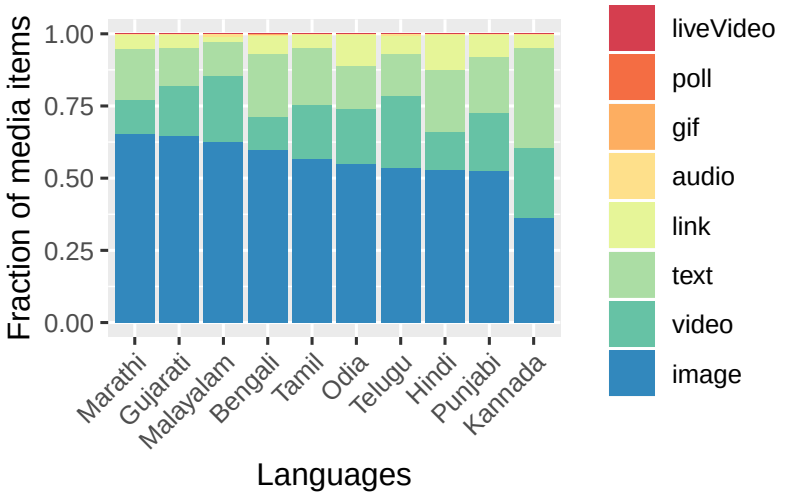


Fig. 5.5 Media count in posts per language. The x-axis is sorted by decreasing count of images in each language.

5.3.2 Media Types

ShareChat supports various media types including images, audio, video, web links and even survey-style polls. Whereas audio (and audio components of video) would be language specific, images are more portable across languages. Figure 5.5 presents the distribution of media types per-post across the different language communities. Overall, images dominate ShareChat with 55% of all posts containing image content. These are then followed by video (20%), text (19%) and links (6%). It is also interesting to contrast this make-up across the languages. Whereas most exhibit similar trends, I find some noticeable outliers. Kannada (and Haryanvi, not shown) have a substantial number of posts containing web links, as compared to other communities. I conjecture that, overall, the heavy reliance on portable cross-language media such as images and web links may aid the sharing of content across languages.

5.3.3 Temporal Patterns of Activity

Given that each language appears to primarily have independent content that is native to its language, I next examine how the activity levels of different languages vary across time. Figure 5.6 presents the time series of posts per day across languages.

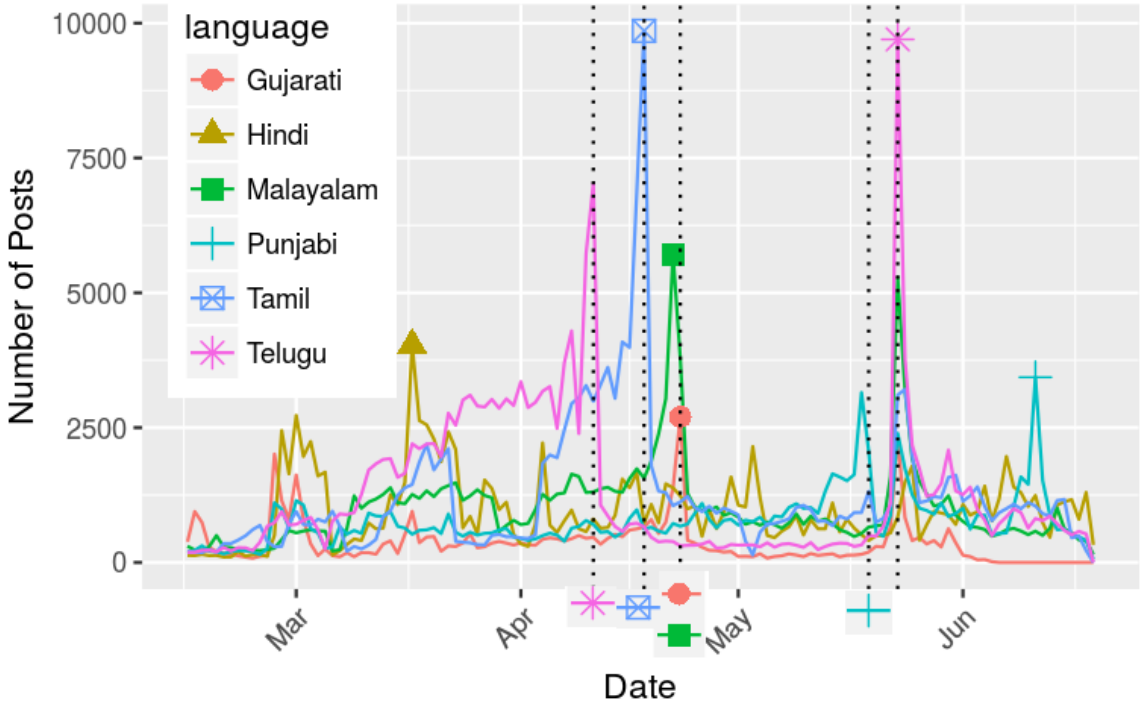


Fig. 5.6 Daily posts plot for languages having maximum peak of more than 2,500 posts per day. Vertical dotted lines are voting days in one of the phases of the multiple-phase Indian election. The right most vertical dotted line is 23 May, when the results of the election were announced across the nation, causing peaks in all languages.

Some synchronisation in activity volumes can be observed across languages, which is likely driven by the electoral activities. Interestingly, however, the peaks of different languages are out of step with each other in many cases. Digging deeper, I find that this is caused by the multiple phases of voting in the Indian Elections, and the peaks correspond to the voting days in the states where those languages are spoken, as marked with the vertical dotted lines (marker below the x-axis shows which state had a voting day corresponding to that peak). For instance, the Voting day for Andhra Pradesh (major Telugu speaking state) is 11th April; Kerala (Malayalam) on 23th April, Punjab (Punjabi) 19th May; Tamil Nadu (Tamil) 18th April. The final peak on 23 May, (corresponding to the election result declaration day) sees a peak for all the languages. Although intuitive, I see that these language trends are *not* agnostic to the underlying events important to their communities.

5.4 Image Spreading Across Languages

As noted earlier, ShareChat is organised into separate language communities, and users must choose a language when registering. Thus, each language community forms its own silo. In this section, I explore to what extent information (more specifically image content) crosses from one language silo to another.

5.4.1 Crossing Languages is Difficult

As noted in §5.2.1, similar images are collected together into clusters using the PDQ algorithm. To test whether users from different languages are exposed to the same image content, I first begin by checking users from how many different languages (as identified by the language set in the profile of the user posting) are represented in each cluster. Figure 5.7 presents the number of image clusters that span multiple language communities. Note that the y-axis is in log-scale, and the vast majority (98%) of clusters are images from a single language community. However, the remaining 2% of clusters transcend language boundaries, with 108 clusters (3258 images in total) crossing five different language communities. This shows that images *can* be an effective communication mechanism, particularly when contrasted with text (where only 0.3% of text-based posts occur in multiple language communities using the same methodology).

I next test if images that cross language boundaries proceed to obtain more "shares", *i.e.,* are they more popular. Recall that shares refer to the act of sharing content via WhatsApp. Figure 5.8 (top) presents, as a box plot, the number of shares per item, based on how many language communities it occurs in. There is a clear trend, whereby multi-lingual content gains more shares. Whereas content that appears in one language community gains a median of 15 shares, this increase to 20 for 4 languages. This appears to indicate that users may expend more energy to translate content that they find to be 'viral' or worthy of sharing widely. This is intuitive as, naturally, images that move across clusters also gain more views. Figure 5.8 (bottom) confirms this, showing that the number of views of images in a cluster

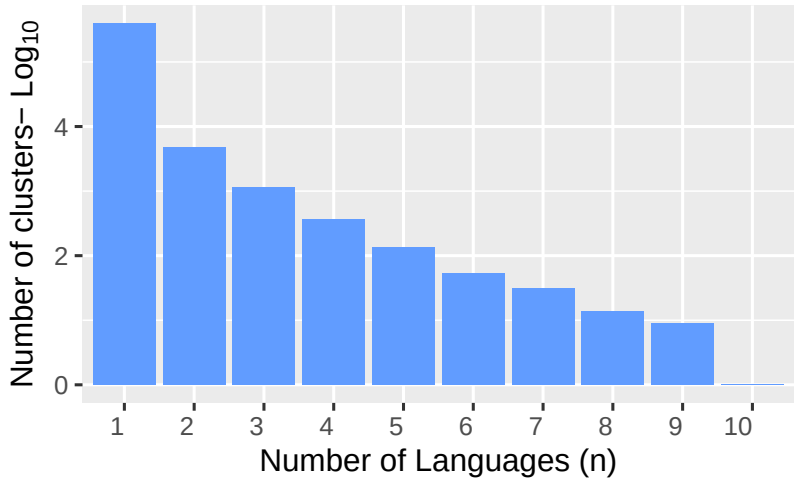


Fig. 5.7 Number of image clusters spanning n languages. Each cluster consists of highly similar images as determined by Facebook PDQ algorithm. *y-axis is log scale.*

increases with the number of languages in that cluster. There is a strong (74%) correlation between number of views on Sharechat and numbers of shares from ShareChat to WhatsApp.

I also noticed the presence of text on the image affects the propensity to share across languages. Recall that 15% of images do not contain any OCR text. I find that this further impacts sharing rates. The average number of shares for images without text is 22 (standard deviation=74) compared to 34 (standard deviation=155) for images with text. A KS-test (one-sided) confirms that this difference is statistically significant ($D = 0.09352, p < 0.001$). For images with OCR text, 80% of OCR text is in the same language as the user's profile language. This leaves around 20% posts that have an OCR text language that is different from the user's profile language. This is important as it confirms that non-native languages can penetrate these communities. This 20% is mostly made-up of lingua franca, *i.e.,* English (11% posts, average share 24), Hindi (8%, average share 34) and Others (1%).

5.4.2 Quantifying Cross Language Interaction

Based on the above observation, I next take a deeper look at which languages co-occur together and consider cross language interactions via both direct text (hashtags), as well as the text contained within images (*i.e.,* OCR text). To measure the extent to which some

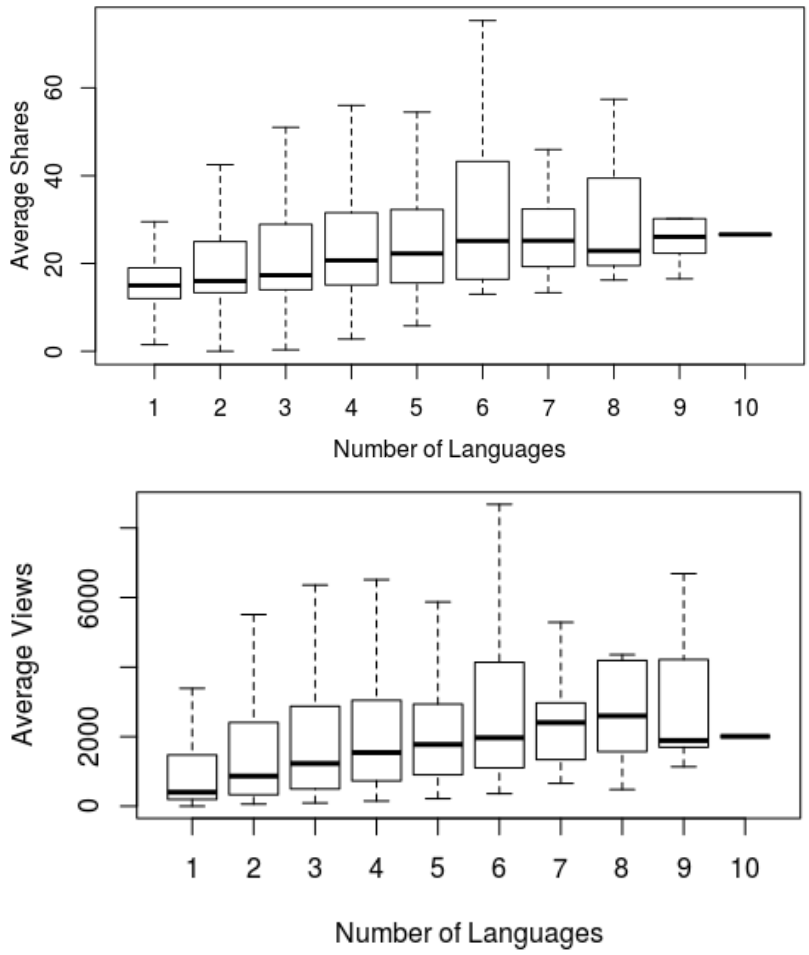


Fig. 5.8 Distribution of the average number of shares (top) and views (bottom) of the image variants represented in a cluster. The median increases with the number of languages where images from that cluster are found.

piece of information may go from one language silo to another, I take all posts from users of a specific language, and detect the languages of tags used on those items using Google's Compact Language Detector 2 [Oom18]. Similarly, for all images posted by users from a specific language, I detect the language of any OCR text embedded in those images.

Figure 5.9 (top) presents the language make-up for tags within each language community or silo, and Figure 5.9 (bottom) shows the language make-up for the OCR text taken from images. I see that in most language communities, the dominant language for both tags and OCR text tends to be the language of that community. However, I *do* observe a number of cases for languages bleeding across community boundaries. Although it is less commonly

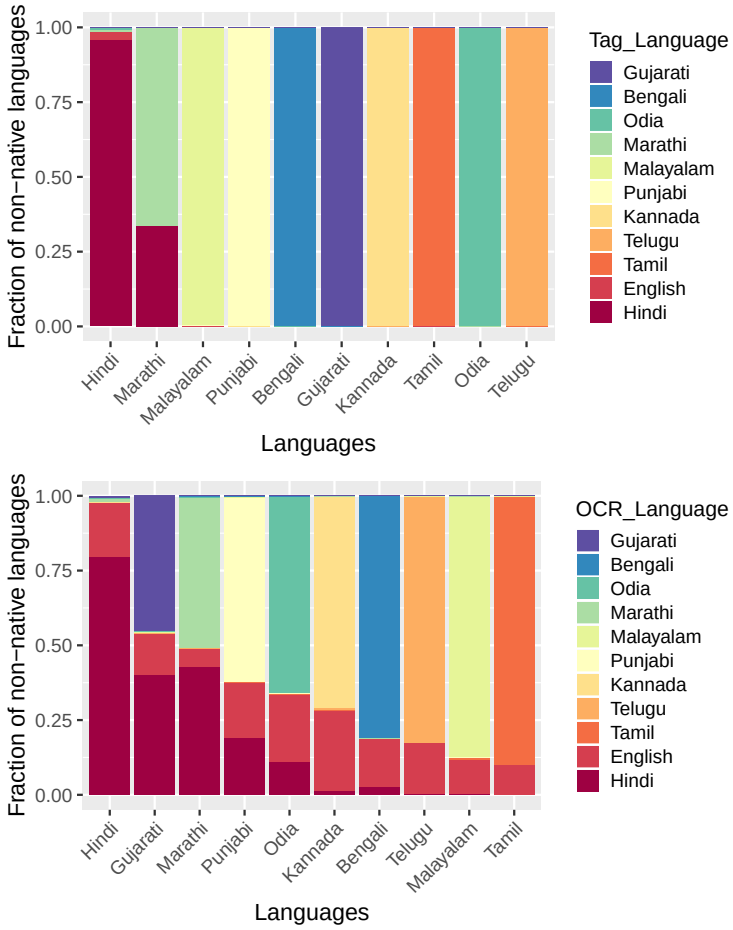


Fig. 5.9 Proportion of non-native languages from (top) hashtags (text) and (bottom) OCR text (images). Languages on x-axis indicate the user profile language and are ordered by descending proportion of English + Hindi, two languages which are used and understood widely across India.

observed in tags, I regularly see images containing text in other languages shared within communities (as discussed in §5.4.1). As the lingua franca, Hindi and English are by far the most likely OCR-recognised languages to be used in other communities. In fact, together Hindi and English make up over half of the image (OCR) text in the Gujarati and Marathi communities. That said, this also extends to other less widely spoken languages too, *e.g.,* the Odia community contains noticeable fractions of tags in English and Marathi, as well as Odia itself. This confirms that it *is* possible to transcend language barriers, although clearly the prominence of the local language shows that this is not trivial.

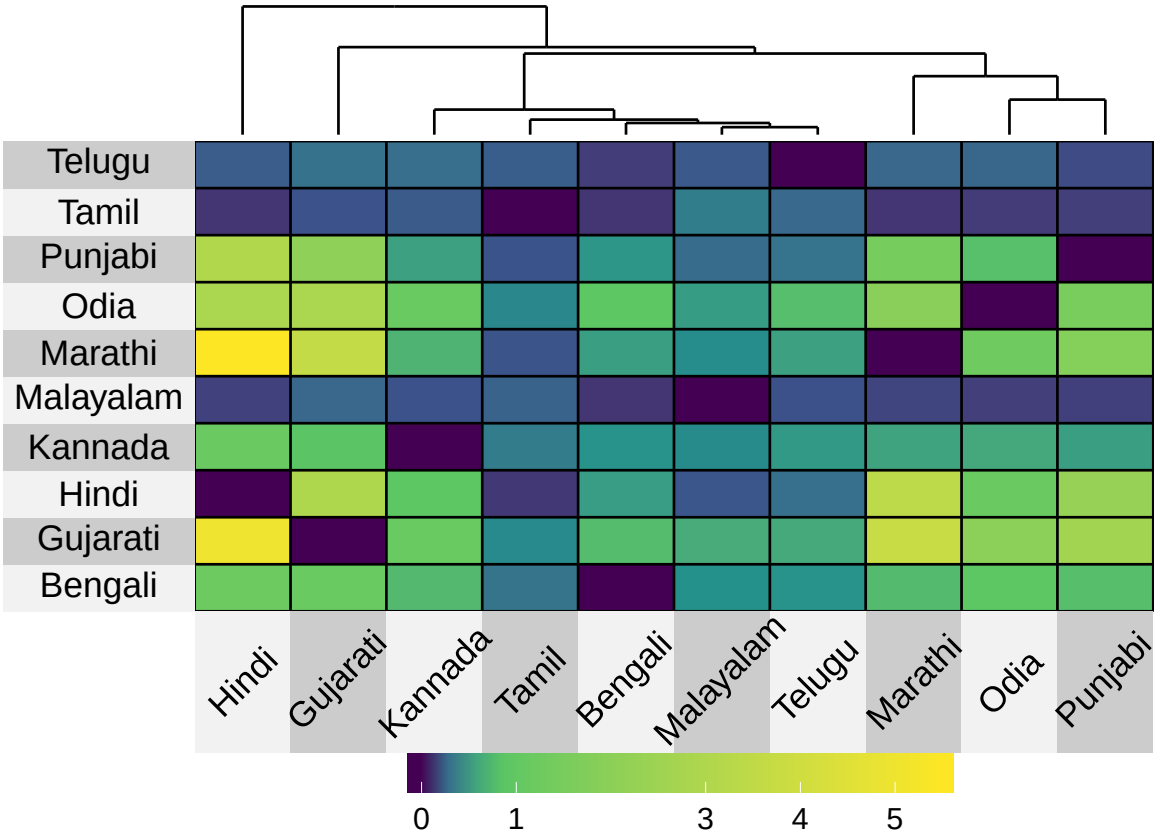


Fig. 5.10 Co-occurrence of languages in image clusters

5.4.3 Drivers of Cross-language Interaction

The previous subsection hints at one possible factor that may lead to the use of a foreign language – the widespread understanding of lingua franca such as Hindi and English. I next explore the extent to which relationships between languages affect whether images are shared across those languages. There are three major language families in India: the Indo-Aryan languages, Dravidian and Munda [Eme56], of which the first two are represented in ShareChat. It is more common to find speakers who can speak both languages of a pair, for languages within the same language family. To get a better understanding of how language communities intersect via image sharing, Figure 5.10 presents a dendrogram to highlight how different language communities tend to share material. I measure the co-occurrence of the same image (or close variants of the same image) in different language communities. Thus,

Table 5.3 Maximum (top 5) and minimum (bottom 5) of overlaps among pairs of languages

Rank	$L1$	$L2$	Overlap
1	Marathi	Hindi	5.48%
2	Gujarati	Hindi	5.01%
3	Gujarati	Marathi	3.72%
4	Hindi	Marathi	3.36%
5	Punjabi	Hindi	3.04%
...	...	...	...
86	Hindi	Tamil	0.13%
87	Tamil	Marathi	0.12%
88	Malayalam	Bengali	0.12%
89	Tamil	Bengali	0.12%
90	Tamil	Hindi	0.12%

I measure the extent to which the same or similar information is conveyed in two different languages, for every possible language pair.

The dendrogram recovers similarities between languages that are geographically or linguistically close to each other (e.g., Dravidian languages such as Tamil, Telugu, Malayalam and Kannada). Table 5.3 shows the fraction of posts of a given language $L2$ that occurs within the community of a language $L1$, focusing on the five pairs of languages with the maximum and minimum overlap. Four of the top five overlaps are observed between Hindi and languages of states where Hindi is widely spoken and understood even if it is not the official state language (Gujarat, Punjab and Maharashtra, which speaks Marathi). The only top overlap where Hindi is not involved is between Marathi and Gujarati, languages of neighbouring states, with a long history of interaction and migration. The bottom of the table shows that language pairs with the least cross-over are those from different language families (e.g., Tamil, a Dravidian language and Marathi, an Indo-Aryan language).

Interestingly, however, the dendrogram also points to close interactions among some pairs of languages that come from very different language families and are spoken in states that are geographically far apart, such as Bengali and Kannada. Manual examination reveals that many of the posts shared between these two languages are memes containing exhortations to vote (Figure 5.11 shows an example). W. Bengal (where Bengali is spoken) and Karanataka (where Kannada is spoken) both went to polls on the same day, which suggests that the

Fig. 5.11 A "go and vote" message shared across Kannada (left), Bengali (right), and other languages.

shared content between these two languages may have resulted from an organised effort to share election-related information more widely.

5.5 Case Study 1: Content Transcending Languages in Multi-lingual Social Media

In this first case study, using the same dataset of Sharechat platform, I look at what *kinds* of content get translated and move across languages. To understand this, I first created a list of the most commonly occurring words in the OCR text of the images (See §5.2.1). Taking all the words which occur more than five times into consideration, I manually code them into 9 categories. I follow a two step approach to come up with this categorisation. First, different authors of the paper coded a small subset of images to come up with a coarse set of categories. These were merged to come up with the final list of 9 categories of information which transcend language boundaries: election slogans, party names, politician names, state names, Kashmir conflict, cricket, India, world and others (such as viral, years, family, share and so forth). Figure 5.12 (top) shows the total number of shares that each of these categories get. This reveals an interesting mix of topics that provide new insights: since I collected

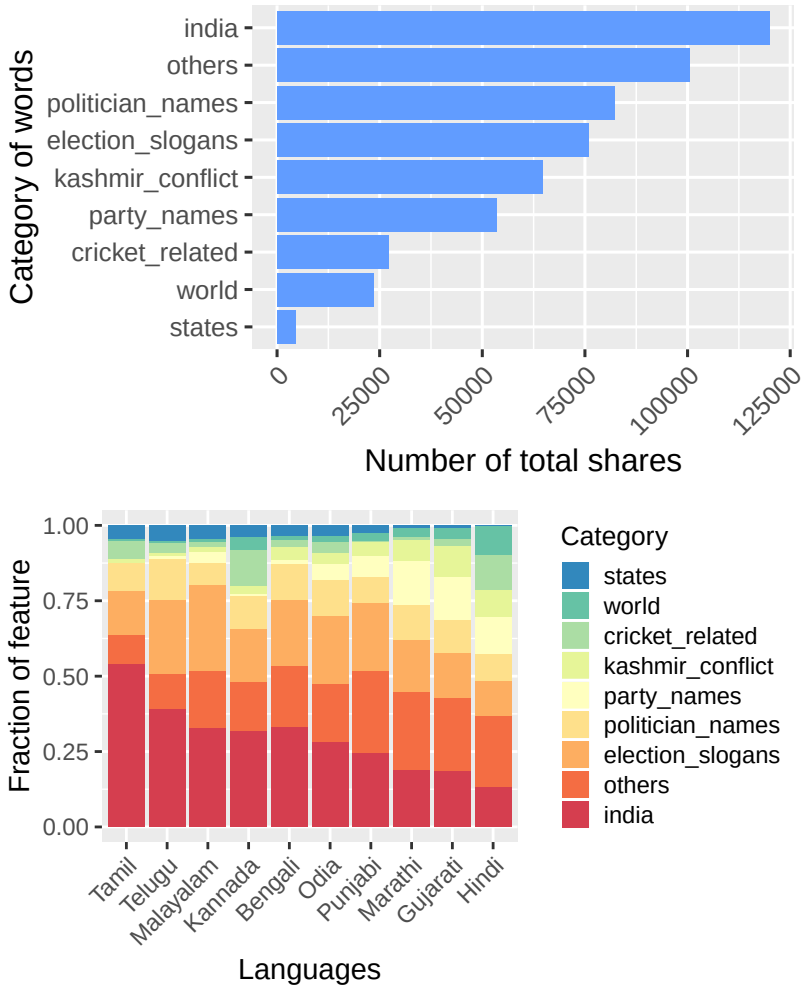


Fig. 5.12 Most popular categories of posts that are shared across language barriers and number of shares (top). Breakdown of topics per language community (bottom).

data related to politics, keywords related to politics such as politician/party names, *etc.* are expected. However, it is interesting to see categories related to issues such as the Kashmir conflict and cricket — these topics are of interest across the nation; possibly an important consideration when deciding which images to translate across languages.

Briefly, Figure 5.12 (bottom) also highlights the presence of these topics across the language communities. I see that trends are broadly similar, yet there are noticeable differences. For instance, "India" makes up over half of the posts in Tamil compared to less than 10% in Hindi. Individual language communities also show different preferences; for instance, Malayalam shares the highest proportion of election slogans. Similarly, Hindi,

Fig. 5.13 Example of images with different messages which portray a political message with different messages in Hindi. Both show Mr. Modi, the incumbent Prime Minister of India at the time of the Election. The caption of the left figure reads "*Friends, I am coming*", whereas the right figure reads "*Friends, I am going*".

Marathi and Gujarati share more party names-related images than other communities, whereas Hindi and Kannada share more cricket images than other languages. Given that Indian states were created based on the languages spoken, cross-language sharing of state-specific issues are negligibly small in all communities.

Previous sections have shown that meme-based images (containing text) that transcend language barriers often benefit from translation. Due to this, I am interested to know whether such translations preserve or alter the semantic meanings. Since many of the most widely translated categories are related to politics or contentious national issues in India, this question acquires an additional dimension of truth and propaganda.

The data has 68 image clusters with multiple languages, consisting of 1080 images. Of these, I am able to fully translate all images within 63 of those clusters.[13] These are made-up of 769 images. To compare the meanings of each piece of text within an image cluster, I allocate each of the images to 2 annotators. Each annotator looks at all of the images in a cluster, as well as the associated OCR text translated into English.[14] The annotators are then responsible for grouping all images within the cluster into semantic groups, based on their OCR texts. This may yield, for example, a single image with two different OCR texts with

[13]The other clusters contain Odiya or Punjabi – languages for which I did not have access to native speakers
[14]The translations are high quality manual translations by native speakers as described in §5.2.1

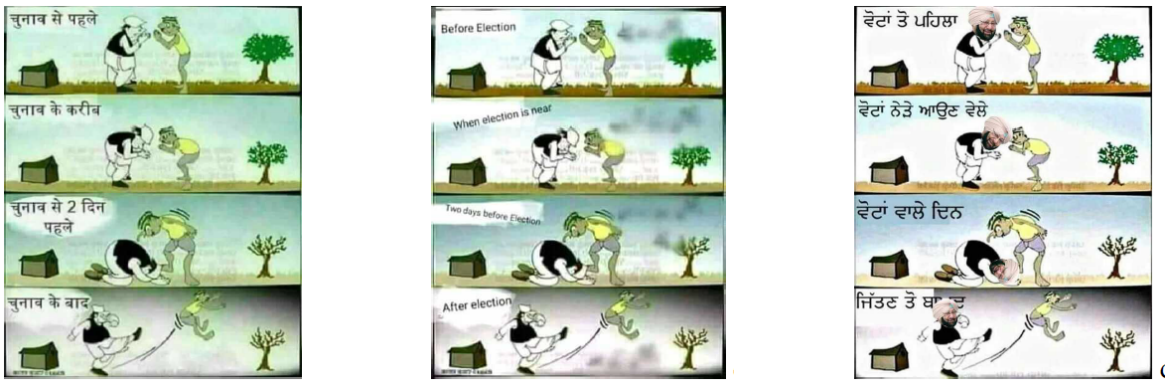


Fig. 5.14 Example of a meme posted in three different languages. These three posts show the same joke about how politicians court citizens, becoming very courteous just before the elections (politician bowing down) but then turn on them after getting elected (bottom image: politician kicking the citizen).

diametrically opposite meanings (e.g., see Figure 5.13). I then say that there are two different "messages" contained in these clusters.

I find that images contained in the majority of clusters have similar messages even when the text is translated into multiple languages (Figure 5.14 shows an example). However, I find a handful of clusters with more than one message: 11 have two messages; often these are memes, but with the "translations" containing distinct messages that have different meanings. Figure 5.15 shows that in all, only 25 clusters have more than one message, and the number of distinct messages in each cluster tends to be small: only 2 clusters have more than 5 messages, and one cluster has 9 different messages. Interestingly, I find that image clusters containing more than one message tends to get more shares (mean 44, median 24) and views (mean 2865, median 1889) than clusters where the images contain only one meaning (mean shares 26, median 20; mean views 2255, median 1329).

Finally, I briefly note that although detailed manual coding and analysis supported by native speakers (§5.2.1) suggests that most of the cases where images have different messages is caused by differences in the *text* embedded in the images, I have also observed a few cases where the images themselves have been changed, creating a new meaning. Figure 5.16 shows an example.

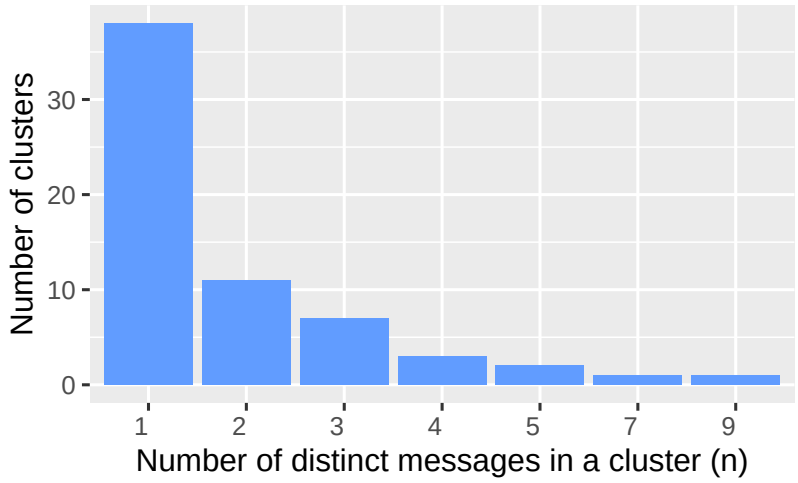


Fig. 5.15 Number of image clusters where OCR texts contain more than *n* distinct messages.

5.6 Case Study 2: Jettisoning Junk Messaging in the Era of End-to-End Encryption

WhatsApp is a popular messaging app used by over a billion users around the globe. Due to this popularity, junk messaging on WhatsApp is an important issue. Despite this, the distribution of junk via WhatsApp remains understudied by researchers, in part because of the end-to-end encryption offered by the platform. This case study addresses this gap by studying spam[15] on a dataset of 2.6 million messages sent to 5,051 public WhatsApp groups in India over 300 days. Public WhatsApp groups are groups where the admins publicly share a link (e.g., via a website or social platform) to join the group. The links allow external observers to join the group and thereby observe the messages exchanged, including spam, although the content remains opaque to the platform itself. Note that these links are intended for people to be able to join and participate. Researchers should be careful while observing the data and should not post or alter the conversation. First, I characterise spam content shared within public groups and find that nearly 1 in 10 messages is spam. I observe a wide selection of topics ranging from job ads to adult content, and find that spammers post both URLs and phone numbers to promote material. Second, I inspect the nature of spammers

[15] I interchangeably use the terms 'spam' and 'junk' to refer unwanted unsoliciated messages.

Fig. 5.16 Example of non-text cluster where political faces are changing. Both images depict politicians as beggars. The left image shows leading politicians from the *opposition* party and the right image replaces those heads with leaders of the *ruling* party.

themselves. I find that spam is often disseminated by groups of phone numbers, and that spam messages are generally shared for longer duration than non-spam messages.

Because of the link-based joining behaviour, public WhatsApp groups typically contain several users who may be strangers to each other, i.e., users who do not have a social connection in the offline world other than via the WhatsApp group. WhatsApp provides strong protection from being contacted by strangers, allowing users to easily block unsolicited messages from those not in their contact list. In contrast, as long as a user is a member of a public WhatsApp group, they cannot avoid messages that strangers may send to that group, regardless of whether the messages sent are germane to the group or not. Thus, spammers can abuse public WhatsApp groups by sending unwanted messages that are irrelevant to the purpose of a group, e.g., links to adult content or sexual services, phony job offers etc. Understanding the nature of spam in public groups is therefore vital for securing WhatsApp and other messaging platforms. In this case study, I focus on India, a country where over 400 of the 460 million people online are on WhatsApp messaging platform. The dataset contains 2.6 million messages from 5,051 public political WhatsApp groups in India (§5.2.2).

Understanding the scale of spam

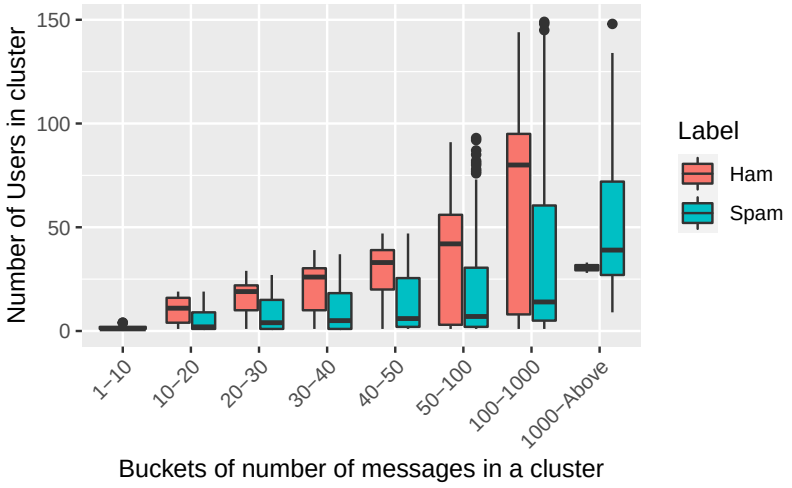
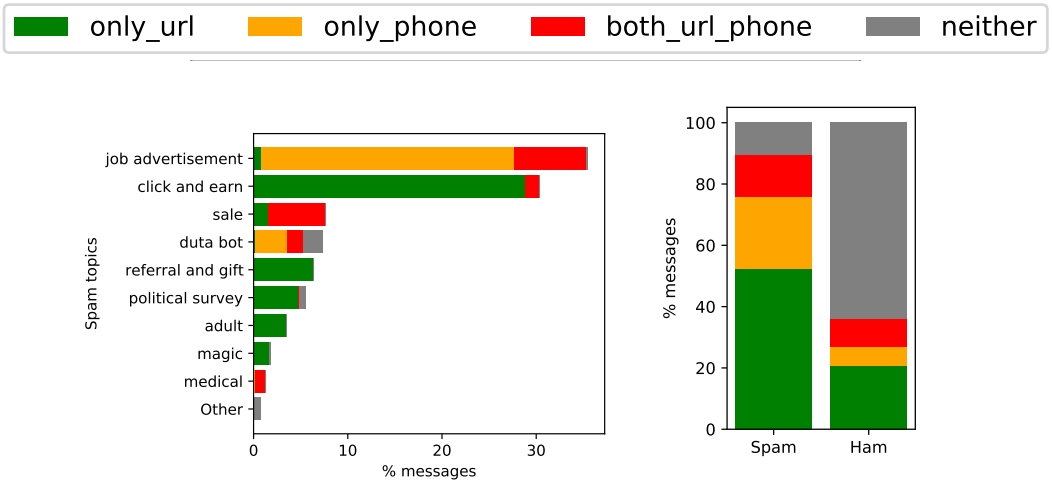


Fig. 5.17 Numbers of users spreading a message, indexed by number of times the message or its close variants are seen.

I first consider two aspects of scale: the number of times a message is posted, and the number of users involved in spreading the message. Each message cluster obtained (see §5.2.2) may contain several messages that are closely related variants. Figure 5.17 groups message clusters into different buckets based on the number of times those messages are found in data, across different WhatsApp groups. I then show how many users were involved in spreading the messages in that cluster.

I find that clusters containing spam messages have larger numbers of messages on average although the median numbers are similar (24 for spam clusters *vs.* 23 for ham clusters; compared to mean 83.6 in spam *vs.* 35 for ham). This indicates a highly skewed distribution of messages per cluster, particularly for spam clusters. For instance, there are over 37 clusters of spam with more than 1000 messages, the largest cluster having over 27K messages. In contrast, only 2 ham clusters have more than 1000 messages and the largest having 2,065 messages. These were videos URLs and associated text related to political speeches.

I also see that popular clusters, which contain messages that are forwarded many times, inevitably involve more users. However, Figure 5.17 shows that for messages in spam and ham clusters of similar sizes (*i.e.,* in the same "bucket" size on the x-axis), the spam clusters



(a) (b)

Fig. 5.18 (a) Junk Topics found in the top 250 clusters of spam (48% of all spam messages); (b) Overall fraction of URLs and phone numbers found in the content of messages

are driven by fewer users. This indicates that spam is disseminated in a more proactive manner by a smaller set of individuals.

Understanding spam content

I next inspect the content of spam posts, including both the topics covered and the sharing of URLs.

To explore the topics discussed within spam, I employ the three annotators to manually examine messages in the top 250 clusters of spam messages. By performing an initial qualitative analysis, I identify 10 core topics. I then ask the annotators to categorise the messages from the top 250 clusters into these 10 topics. The top 250 clusters comprise 61% (181K messages) of the total number of spam messages (295K) in this dataset.

Each message cluster is examined by at least one annotator and the category label applied was then checked by a second annotator. At least one of the two annotators was a native speaker of one of the four languages considered (Hindi, Telugu, Tamil or English). All differences of opinion were resolved by a bilateral discussion and I finally obtained 100% inter-agreement between annotators.

In addition to strategies followed by traditional social media spam — using click-bait or other techniques to take users outside the platform, WhatsApp spam also makes the use

of phone numbers. Figure 5.18a captures the relative frequencies of the topics, as well as the frequency of URLs and phone numbers within the messages. Each of the 10 topics is described below:

Job Advertisements. Comprising nearly 35% of the annotated spam, the most widespread spam is advertisements for jobs. Nearly all (99.6%) of job advertisements provide a contact phone number, or a phone number as well as a URL. Of course, these may be genuine job advertisements, although they are off-topic spam for political WhatsApp groups. Though I was not able to identify if these were genuine job ads, the template structure of these ads makes it believe these are spam and could involve a scam. Interestingly, over 97% of spam in Telugu language forums consist of job advertisements.

Click and Earn. These comprise 30% of spam messages, and ask users to click on a URL, promising a reward. 99% of these messages contain a URL, but no phone number.

Sales. These constitute 7.6% of spam and offer items for sale. 79% of these messages contain URLs and a phone number, and could be genuine items for sale.

Duta Bot. These are (benign) spam messages (7.3%) sent by a news bot service called Duta Bot [DUT18] comprising regular news or sports updates.

Referral and gifts. These spam messages (6.3%) offer a gift in return for referrals of users to an online service subscription, and consist mostly of a URL to click.

Political Survey. These (5.6%) mostly contain URLs that invite users to participate in political surveys. These are mostly benign and also partly on-topic as the WhatsApp groups I consider are political.

Adult. These (3.4%) mostly contain URLs that lead to adult websites or offer adult sex-related services.

Magic. These (1.8%) messages contain text which asks user to forward a message to experience something supernatural, *e.g.,* "Forward and see magic: your phone battery will get charged to 100%".

Medical. These (1.2%) messages offer treatment for common and sometimes embarrassing ailments, *e.g.,* "ayurvedic treatment for piles".

Other. Approximately 1% cannot be categorised into any of the above groups and consist of spams such as "daily event update".

Next I look at the *users* who have produced spam in terms of their temporal patterns and group membership.

Operational definition of spammers

My methodology identifies spam messages by their content rather than spammers directly. Some spam messages may be inadvertently posted by enthusiastic or naïve users who do not realise it is spam. As expected, this follows a bimodal pattern (cf Appendix 7.3), with spammers on one end (nearly 100% of their messages are spam) and non-spammers ("*hammers*") at the other end (with almost no spam messages). This suggests that users with a spam fraction beyond any reasonable threshold such as $f = 50\%$ will capture all intentional spammers. In this section, I adopt an operational definition of spammers as any user who has posted more than $f = 50\%$ messages that my methodology identifies as spam (my results are robust to other similar thresholds). Using this methodology, I identify 17.6K users as spammers and 32.9K as non-spammers who share a total of 239K and 1.3M messages, respectively.

Longevity patterns of spammers

I define a cluster of lexically close spam messages (specifically, messages that map to the same LSH cluster in my pre-processing) as a *spam campaign* and ask whether there are longer time-scale patterns, or focused campaigns in spreading the same messages. I term any day when at least 10 messages relating to a campaign are posted as an *active day* for the campaign. I start by inspecting the lifetime of spam *vs.* ham messages, in terms of the difference between their first and last occurrence, shown in Figure 5.19. I see that spam

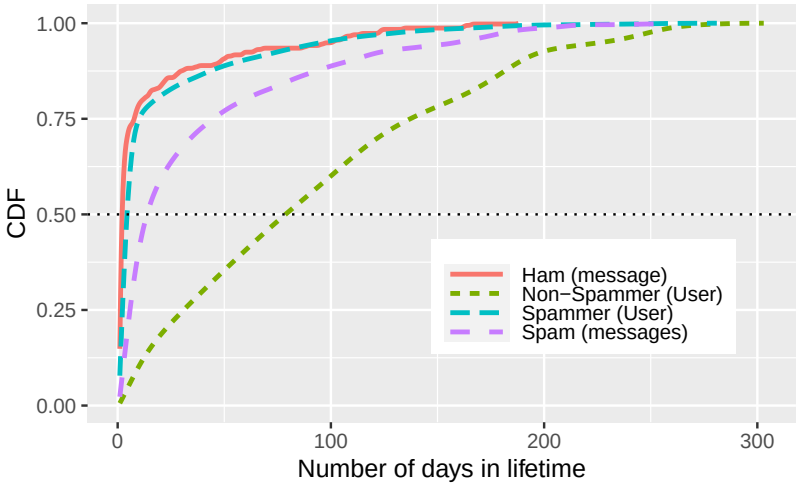


Fig. 5.19 Cumulative Distribution Function (CDF) of lifetimes of spam and ham message clusters and users.

campaigns have consistently longer lifetimes than ham messages. The median campaign duration for spam is 29 days compared to 12 days for ham. This means the same messages are sent over a longer period, potentially allowing better capture of attention [AS90, S$^+$90]. Interestingly, I see the opposite trend when computing the lifetime of accounts. Here, I compute the lifetime of a user as the difference between their first and last post. As shown in Figure 5.19, non-spam users have substantially longer lifetimes than their spammer counterparts.

The above leads us to explore the daily characteristics of these spam campaigns. Figure 5.20 presents the number of messages sent per-day for the top 15 spam campaigns (ranked by number of times the message is posted). These campaigns occurs across multiple days and are highly focused, with aggressive peaks on a small number of days.

Joining and leaving groups

I next examine how users (spammers and others) join and leave different groups. Note that users may be added by another user, by an admin or they can join with an invite link. Figure 5.21 presents the fraction of spammers *vs.* non-spammers who join using these techniques. A clear difference exists: it is much more common for spammers to join WhatsApp groups via a link. Spammers also comprise a disproportionate fraction of users who are 'removed'

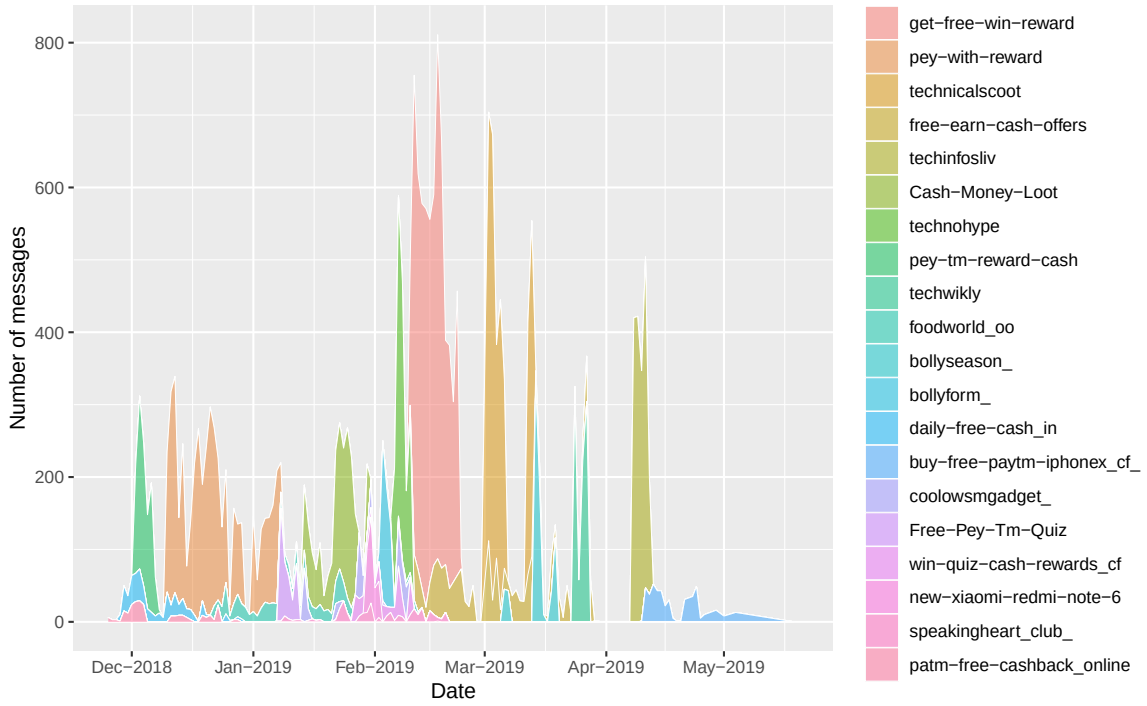


Fig. 5.20 Number of times a day that a spam message is posted, for 15 exemplar campaigns.

Table 5.4 Relation between leaving and joining methods for spammers (equivalent numbers for hammers in parenthesis).

Actions	**Joined via link**	**Added**	**Added by admin**
Left	75%(62%)	17%(20%)	8%(18%)
Number changed	74%(37%)	12%(28%)	14%(35%)
Removed	80%(48%)	5%(18%)	15%(34%)

from a group (this action is usually undertaken by admins when users violate the norms of a group).

Table 5.4 examines more closely how users leave groups and how these actions are related to the method they used to join. Users who leave the group by any method are overwhelmingly likely to have joined via an invite link, although this is noticeably higher for spammers than hammers. This suggests that such users are less involved in the group. Note that users who have been added by a group admin are the least likely to have left the group, but a fair number of those added by an admin are also later forcibly ejected.

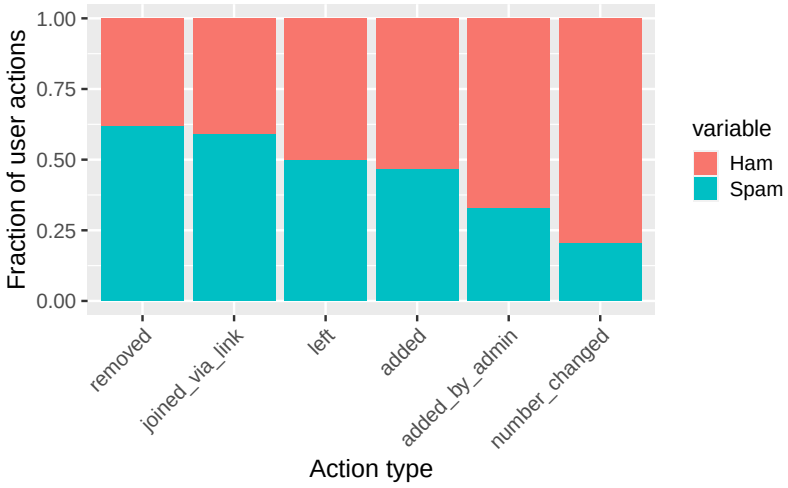


Fig. 5.21 Relative proportions of join and leave actions.

I further inspect what users *do* after joining. I find that after joining via a link, nearly 40% of spammers (30% of non-spammers) post URLs. 18% (5%) post messages with a phone number, and 21% (21%) simply leave. Again for spammers, after posting a spam message with a URL, 86% of the time the next action is to post another spam message with a URL. I also check the actions of spammers immediately before they are removed from a group. I find that 54% of the time they posted a spam message with a URL, and 19% of times they post a spam messages with a phone number in it. In total, 73% of user removals by admins are immediately after the user posts a spam message.

5.7 Conclusion

This chapter investigated the role of language in the dissemination of political content within the Indian context, through the lens of ShareChat. I began by asking three sets of research questions.

First, *can image content transcend language silos and, if so, how often does this happen and amongst which languages?* I find that the vast majority of image clusters remain within a single language silo. However, this leaves in excess of 33k images crossing boundaries. I find that geography and language connections play a role: when content crosses language

boundaries, it does so into languages which belong to neighbouring states, or which are related linguistically.

Second, *what kinds of images are most successful in transcending language silos, and do their semantic meanings mutate?* I certainly observe certain topics that are more effective at gaining cross language interest. As anticipated, I find that images containing text related to national Indian politics cross languages more often. But I further observe other topics of pan-Indian interest (*e.g.,* cricket) also gain traction among a diverse set of languages. By clustering images based on perceptual similarity, and manually verifying their semantic meanings with annotators, I found that 25 clusters of images had text which changed meaning as they crossed language barriers.

Third, *what are the emerging challenges which platforms need to consider for tackling misbehaviour and maintaining political neutrality?* I find that spam is commonplace on platforms like WhatsApp and spammers tend to post across a large number of groups and languages. Spammers also exhibit interesting patterns of leaving and re-joining groups multiple times, to avoid being removed by admins. I further find evidence of spammers coordinating campaigns — spreading the same spam message (or close variants) over a small number of 'active' days. These strategies may help improve the visibility of spam by providing a longer 'shelf life' in the recent messages.

CHAPTER 6

REFLECTIONS, FUTURE DIRECTIONS AND CONCLUSION

The secret of getting ahead is getting started.

Mark Twain

This work is an initial attempt to provide evidence-based support for policy and regulation that can safeguard the nationally important discourse between Politicians and citizens in the digital era. To this end, I study political engagement model that involve bi-directional online interactions among of politicians, citizens and digital mass media platforms at scale. In each form of political engagement I establish methodologies, characterise datasets, and provide various case-studies as examples. The work done as a part of this research helps in understanding the current and emerging state of digital citizen engagement involving multiple countries and platforms. Finally, I conclude this dissertation with reflections of my contributions, limitations which I encountered and future research directions of this research work.

6.1 Reflections and Contributions

In Chapter 3, I first collected and studied a dataset of covering months of conversations between MPs and citizens[1]. My data captured entire threads of conversation by taking advantage of new changes to the Twitter API. I then present online harms such as hate

[1] https://nms.kcl.ac.uk/netsys/datasets/tweeting-mps/

speech from citizens to MP and characterise along demographics, party affiliation and other information. I then examined the prevalence of hate among different groups, finding evidence that there was an increased amount of hate towards MPs from ethnic minorities, but contrary to studies in other contexts [FS19, EZ21], I find that male and female MPs received equal amounts of hate. I also showed that a significant proportion of hate comes from across party lines, with MPs of the Conservative Party (the party which was in Government during the period of my study) receiving more hate than other parties. I also identified a "pile on" effect whereby MPs who are in the news and are already getting a high volume of tweets for one reason or another tend to receive more hate.

The Draft Online Safety Bill published recently by the UK Government [Par21] would impose various duties on providers of online user-to-user services in respect of harmful content while at the same time protecting users' rights to freedom of expression and privacy. I hope that my findings can contribute to the development and implementation of this or similar regimes in other countries by helping to develop more accurate and proportionate computational methods for identifying hateful content. I also hope this dataset will be useful for qualitative research from humanities and social sciences. It may provide political scientists with important insights when looking into the phenomenon of online hate, how it emerges and what effects it has. Legal and policy researchers can look into examples such as the dataset I have curated in order to formulate evidence-backed regulatory responses to this new problem.

In Chapter 4, I study news media ecosystem and helped in raising awareness and the need for appropriate privacy laws, restrictions and reducing polarisation. I find that a great majority of news websites are being covered in tracking by individual companies such as Google's *doubleclick.net*, but they also track personas visiting these news sites. Potential privacy laws should pay particular attention to these "monopoly", data aggregation practices, since such companies can be the target of legal or illegal querying for information on specific individual's preferences or large user audiences, by governmental or other advertising

agencies, for political or other marketing purposes. I released all data, code and methods for reproducibility and insensibility purposes[2].

Also, in the case study with Wikipedia and UK Politicians profiles, I observed signals of partisanship among edits and domains of citations on papers using the collected dataset[3]. Many editors contribute to the pages of MPs of one party, and I also see communities of editors who collectively focus on each party. I also see a different distribution of citation ideology scores between pages of MPs from the two main UK parties, with Labour MP pages drawing on sources with a slightly higher degree of ideological bias.

Finally in Chapter 5, I study media and citizens as a first study of multilingual social media and peer-to-peer messaging service at scale. I find that for users of internet in developing nations, image sharing is prevalent and junk is a commonplace even in messaging services like WhatsApp. Content posters and junk senders tend to post across a large number of groups and also exhibit interesting patterns specific to language and user groups. For example the junk sender reuse the same content and exhibits leaving and re-joining patterns in groups multiple times, to avoid being removed by admins.

My results have clear implications for understanding content sharing and junk detection in multi-lingual setting. The dataset is also released for research purpose with anonymised labels [4].

6.2 Limitations

A further research is needed before these findings can be generalised beyond the dataset I use. My dataset (and any such dataset) has to be specific to a particular time period and geography. Although, I believe there are likely common characteristics in hateful speech across borders (that are also reflected in my dataset) it could be the case that my dataset only reflects political discussions in the UK; it could also be that the tone and character of such discussions on other platforms or other time periods may be different.

[2] https://nms.kcl.ac.uk/netsys/datasets/partisan-tracking/
[3] https://nms.kcl.ac.uk/netsys/datasets/wikipedia-mps/
[4] https://nms.kcl.ac.uk/netsys/datasets/share-chat/, https://nms.kcl.ac.uk/netsys/datasets/whats-app/

Indeed, there are some limitations of my framework, which I plan to overcome in the near future. For example, data protection laws do not allow cross-border tracking, which results in logging 33% fewer trackers when the crawling is performed from websites under EU's jurisdiction, as compared to traffic logged within US. Privacy laws need to promote better transparency from third-parties towards their tracking objects, and the reasons for doing so more intensely in specific sub-topical pages (*e.g.,* COVID-19-related products advertised in COVID-19-related topical pages, after intense tracking of users in said pages). Another area to explore is the variability in tracking of personas, and its relation with the ad-economies and end-user experience. For example, if a website is visited on two consecutive days, tracking metrics will most likely be different. Also, there is a wide-range of different, less or more pronounced and compound personas that someone can emulate with my framework. Clearly, every persona comes with costs of finding appropriate websites to build it, and additional experiments to make sure it triggers network traffic realistic for an online user of that persona. In fact, less pronounced personas may not be able to collect appropriate and conclusive data of differential tracking. In the future, I aim to build better tools to understand these problems. I also plan to further develop my framework to look into the ad-ecosystem on fringe news websites, and replicate my work for multiple countries to discover user differential tracking and personal data leaks. The findings may be sensitive to the choice of day, time and place of crawling due to the dynamic ecosystem of tracking.

I note that the tools used above are not robust against major changes in posting strategies (*e.g.,* use of QR codes rather than URLs or meme generation with same semantic but entirely different aesthetics) or changes to how WhatsApp operates (*e.g.,* disabling joining via links). However, I emphasise that these feature set can be expanded, and my models can easily be retrained in response to such adaptations.

6.3 Future Directions

Interaction between Politicians and Citizens. In future, there is additional work to be done. For instance, despite its current widespread use, there still remain some concerns about

how representative Twitter is, as a (or the main) platform for digital citizen engagement. Furthermore, I need to go beyond the current observational study, to conclusively understand whether Twitter engagement is helping MPs in their day-to-day duties or if it merely adds to their burdens. Other questions – such as whether Twitter remains a place for "empty" conversations, or whether actual Government or Parliamentary activity result from these online discussions – need closer scrutiny. Case studies such as the creation of the Assaults on Emergency Workers (Offences) Act 2018 point to ways in which Parliamentary activity can be facilitated or directed through Twitter and other online means, but these are early examples, and there may be other mechanisms that become more commonplace in the near future. In this context, it may be interesting to compare with other more formal routes, including e-petitions, which can get discussed in parliament if sufficient numbers of citizens declare interest in an issue. There are also avenues of exploring more using the data in this research. For example a future work can use the collection of data on MPs not on Twitter, it remains unclear how MPs which are not on Twitter manage (if any) their attention.

Online hate is an important problem as it may be dissuading targeted demographics from fully participating in the national political spheres of several countries [Sco19, Par19]. At the same time online presence is regarded as essential in politics [JL11], so abstaining from this sphere is not an option. Reducing the incidence of online hate is therefore important to prevent representative democracies from becoming less representative of their populations. The phenomenon's deleterious effects on democratic processes has triggered intense policy dialogue, law reform efforts [Com20b] and proposals to create new duties for platforms that may be hosting harmful content [Par21]. More and more research has been emerging on the challenges to managing and countering online hate from a plethora of disciplinary perspectives. I argue that it's time to lay the groundwork for meaningful communication and cross-fertilisation of these perspectives.

Interaction between Politicians and Digital Mass Media Platforms. My dataset was primarily focused on websites using English language. Multilingual online users consist of a large portion in online political engagement [AGJ$^+$20]. However, the diversity of languages raises the question: Are the patterns of tracking similar or different among say regional News

websites? Templates derived from the reference points and cases in Western settings can only partially explain the underlying political dynamics in other countries. Political parties in some countries typically defy linear binaries of Left and Right. In such a context, the coverage bias and media effects are variable and are contingent upon subject, personalities, and circumstances. While the categorisations herein of "Left" and "Right" have been used as a heuristic tool, future research should dive into the contextual specifics of multiple political lines, and offer analysis with finer granularity of the political spectrum. This evidence of personalisation and polarisation therefore might not necessarily imply political bias, but further research in this direction should investigate the neutrality of the content coverage in those articles.

Interaction between Citizens and Digital Mass Media Platforms. A future work can focus more closely on the semantic nature of images, including the characteristics that lead to more popular posts, as well as posts that can overcome language barriers. Preliminary analysis of the ShareChat content has revealed the presence of "fake news", and has shown how it tends to gain higher share rates than mainstream content. Thus, another line of investigation is to trace the origins of such content and understand how it may link to more targeted political activities [BJBS18, AJP+20, Sta17]. I have observed that campaigns tend to encompass multiple languages and groups and, thus, a more rigorous network-based analysis of the dissemination could be useful. I also wish to inspect further the nature of the junk being sent. I have already performed an analysis of URLs, however, I plan to collect further data on the websites these URLs host. I posit this will further support my detection work and hope that this could further feed into junk mitigation efforts on WhatsApp . To aid reproducibility, my annotated dataset (c.f. §5.2.2) is publicly available for researchers[5]. However, as WhatsApp uses end-to-end encryption, such information cannot be accessed by the platform. I will therefore propose techniques that can be used by the end device or the platform (centrally) to identify junk senders, whilst still respecting end-to-end encryption guarantees. For example, a key indicator of junk is the presence of particular URLs and phone numbers.

[5]More information is available at: http://tiny.cc/netsys-whats-app

CHAPTER 7

APPENDIX

7.1 Hate Speech Datasets used for Classifiers

Hate speech datasets used for training classifiers are listed in Table 7.1.

Table 7.1 Dataset details for training classifiers.

Datset	Source	Training Size	%Hate: Davidson model	%Hate: Wulcyzn model
Davidson [DWMW17]	Twitter	25k	1.3	86
Founta [FDC+18]	Twitter	80k	13.5	2.4
Gilbert [dGPGPC18]	Stormfront	10k	11.4	0.04
Jing-Gab [QBL+19]	Gab	34k	0.2	75
Jing-Reddit [QBL+19]	Reddit	22k	22.3	0.09
Kaggle [Kag18]	Kaggle	223k	78.6	2.8
Wazeem [WH16]	Twitter	17k	20.2	2.5
Wulcyzn [WTD17]	Wikipedia	115k	41.5	3.1
Zampieri [ZMN+19]	Twitter	14k	37.6	7.7

7.2 Examples of Junk Messages

Figure 7.1 shows several examples of junk messages.

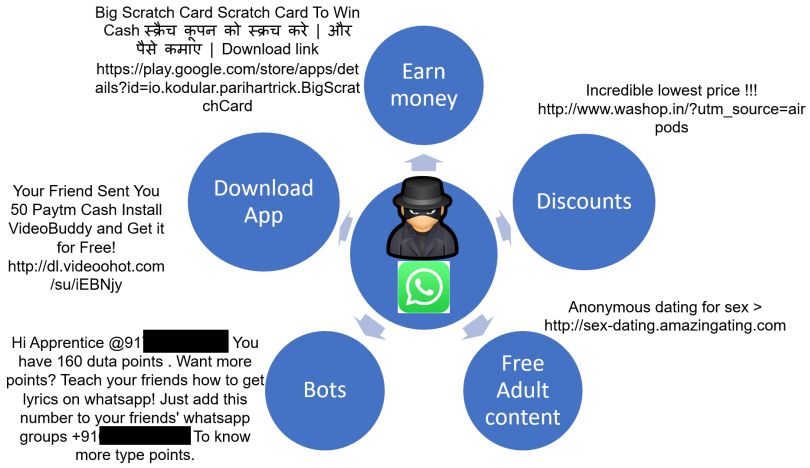


Fig. 7.1 Some top examples of junk messages.

7.3 Junk and Non-junk Senders' Bi-modal Distribution

Figure 7.2 plots the fraction, f, of messages by a user that are marked as junk with respect to total messages. As expected, this follows a bi-modal pattern, with junk senders on one end (nearly 100% of their messages are junk) and non-junk senders at the other end (with almost no junk messages). This suggests that users with a junk fraction beyond any reasonable threshold such as $f = 50\%$ will capture all intentional junk senders. In this study, I adopt an operational definition of junk senders (jettison) as any user who has posted more than $f = 50\%$ messages that my methodology identifies as junk (my results are robust to other similar thresholds).

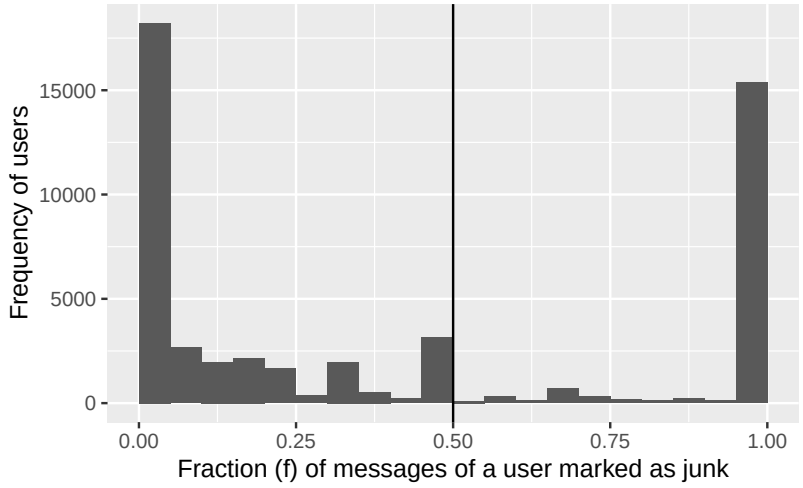


Fig. 7.2 Fraction (f) of messages of a user marked as junk.

REFERENCES

[Ade19] Adext Corp. How programmatic advertising has changed the online advertising landscape. https://blog.adext.com/programmatic-ads-changed-online-advertising/, 2019.

[AEE+14] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the SIGSAC Conference on Computer and Communications Security*, pages 674–689. ACM, 2014.

[AGJ+20] Pushkal Agarwal, Kiran Garimella, Sagar Joglekar, Nishanth Sastry, and Gareth Tyson. Characterising user content on a multi-lingual social network. In *Proceedings of the AAAI ICWSM 2020*, volume 14, pages 2–11, 2020.

[AHA+21] Pushkal Agarwal, Oliver Hawkins, Margarita Amaxopoulou, Noel Dempsey, Nishanth Sastry, and Edward Wood. Hate speech in political discourse: A case study of uk mps on twitter. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 5–16, 2021.

[AJP+20] Pushkal Agarwal, Sagar Joglekar, Panagiotis Papadopoulos, Nishanth Sastry, and Nicolas Kourtellis. Stop tracking me bro! differential tracking of user demographics on hyper-partisan websites. In *Proceedings of ACM WWW*, pages 1479–1490, 2020.

[AK13] Seth Schoen Adi Kamdar, Rainey Reitman. Nsa turns cookies (and more) into surveillance beacons. https://www.eff.org/deeplinks/2013/12/nsa-turns-cookies-and-more-surveillance-beacons, 2013.

[API21] Google's Perspective API. Using machine learning to reduce toxicity online. https://perspectiveapi.com/, 2021.

[ARI+22] Pushkal Agarwal, Aravindh Raman, Damiola Ibosiola, Nishanth Sastry, Gareth Tyson, and Kiran Garimella. Jettisoning junk messaging in the era of end-to-end encryption: A case study of whatsapp. In *Proceedings of the ACM Web Conference 2022*, pages 2582–2591, 2022.

[ARS+20] Pushkal Agarwal, Miriam Redi, Nishanth Sastry, Edward Wood, and Andrew Blick. Wikipedia and westminster: Quality and dynamics of wikipedia pages about uk politicians. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 161–166, 2020.

[AS90] Punam Anand and Brian Sternthal. Ease of message processing as a moderator of repetition effects in advertising. *J. Mktng. Rsrch.*, 1990.

[AS19] Venkat Ananth and Samidha Sharma. On instagram, in india, it's sex for sale. The Economic Times, 2019.

[ASW19] Pushkal Agarwal, Nishanth Sastry, and Edward Wood. Tweeting mps: Digital engagement between citizens and members of parliament in the uk. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 26–37, 2019.

[AU18] Katrin Auel and Resul Umit. Explaining mps communication to their constituents: Evidence from the uk house of commons. *The British Journal of Politics and International Relations*, 2018.

[AVA+21] Vibhor Agarwal, Yash Vekaria, Pushkal Agarwal, Sangeeta Mahapatra, Shounak Set, Sakthi Balan Muthiah, Nishanth Sastry, and Nicolas Kourtellis. Under the spotlight: Web tracking in indian partisan news websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2021.

[B+19] Shakuntala Banaji et al. Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india. *Dept. of Media and Communications, LSE*, 2019.

[B+21] Cedric Burton et al. European commission proposes new rules for digital platforms. http://bit.ly/jdsupra2021, 2021.

[Bar15] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. In *Political analysis*. Cambridge University Press, 2015.

[BARS22] Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K Singh. Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2022.

[BARW16] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. Tracing information flows between ad exchanges using retargeted ads. In *25th USENIX Security Symposium*, 2016.

[BB19] Victor S Bursztyn and Larry Birnbaum. Thousands of small, constant rallies: A large-scale analysis of partisan whatsapp groups. In *ASONAM*. IEEE, 2019.

[BC01] Patrick Butler and Neil Collins. Payment on delivery-recognising constituency service as political marketing. *European Journal of Marketing*, 2001.

[Ber18] Leonid Bershidsky. Yes, russia abused facebook. but did it work? https://www.bloomberg.com/opinion/articles/2018-12-18/yes-russia-abused-facebook-but-did-it-work, 2018.

[BHC+12] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012.

[BHG+20] Jonathan Bright, Scott Hale, Bharath Ganesh, Andrew Bulovsky, Helen Margetts, and Phil Howard. Does campaigning on social media make a difference? evidence from candidate use of twitter during the 2015 and 2017 uk elections. *Communication Research*, 47(7):988–1009, 2020.

[BHK16] Andrew O Ballard, D Sunshine Hillygus, and Tobias Konitzer. Campaigning online: Web display ads in the 2012 presidential campaign. *PS: Political Science & Politics*, 2016.

[BHMDW15] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. Content and network dynamics behind egyptian political polarization on twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015.

[BJBS18] Shweta Bhatt, Sagar Joglekar, Shehar Bano, and Nishanth Sastry. Illuminating an ecosystem of partisan websites. In *Companion of The Web Conference*, pages 545–554, 2018.

[BJN+15] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[Bla20] Marie Black. Most common whatsapp scams, 2020. bit.ly/spam-2020.

[Ble12] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[Bol16] John Boland. The role of the media in the polarisation of british politics. Open Democracy, Aug 2016.

[BTGM04] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 2004.

[Bur74] Edmund Burke. Speech to the electors at bristol at the conclusion of the poll, Nov 1774. Available at https://bit.ly/2jzHV9T.

[BV08] Auksė Balčytienė and Aušra Vinciūnienė. Political communication culture with a european touch: A view from brussels. *Sociologija. Mintis ir veiksmas*, pages 71–85, 2008.

[BV11] Kees Brants and Katrin Voltmer. *Political communication in postmodern democracy: Challenging the primacy of politics*. Springer, 2011.

[CKB+17] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, 2017.

[CMC+15] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. I always feel like somebody's watching me: Measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '15, 2015.

[Col05a] Stephen Coleman. Blogs and the new politics of listening. *The Political Quarterly*, 2005.

[Col05b] Stephen Coleman. New mediation and direct representation: Reconceptualizing representation in the digital age. *New Media & Society*, 7(2):177–198, 2005.

[Com98] Federal Trade Commission. Children's online privacy protection rule (coppa), 1998. https://www.ftc.gov/.

[Com20a] Law Commission. Harmful online communications: The criminal offences. http://bit.ly/harmful2020, 2020.

[Com20b] Law Commission. Hate crime – consultation paper summary. http://bit.ly/hate-crime2020, 2020.

[Cor08] Gordon V Cormack. *Email spam filtering: A systematic review*. Now Publishers Inc, 2008.

[Cow00] Frank Alan Cowell. Measurement of inequality. *Handbook of income distribution*, 1:87–166, 2000.

[Cow02] Philip Cowley. *Revolts and rebellions: Parliamentary voting under Blair*. Politico's, 2002.

[CRF+11] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[CS12] Rafael Cappelletti and Nishanth Sastry. Iarank: Ranking users on twitter in near real-time, based on their information amplification potential. In *2012 International Conference on Social Informatics*. IEEE, 2012.

[dFM+19] de Freitas Melo et al. Can whatsapp counter misinformation by limiting message forwarding? In *Conf. Compl. Netw. and Appl.*, 2019.

[dGPGPC18] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, 2018.

[Dis13] Disconnect, Inc. Disconnectlist - disconnect tracking protection project, 2013. Available at github.com/disconnectme/disconnect-tracking-protection, accessed on 24 June 2020.

[dLPS+17] Agata de Latour, Nina Perger, Ron Salag, Claudio Tocchi, and Paloma Viejo Otero. *We can!: Taking Action against Hate Speech through Counter and Alternative Narratives (revised edition)*. Council of Europe, 2017.

[Dug17] Maeve Duggan. Online harassment 2017. *Pew Research Center*, 2017.

[DUT18] DUTA. Duta.in: Bringing the internet to the next billion, 2018. https://duta.in/index.php.

[DVF+16] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016.

[DWMW17] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2017.

[Eme56] Murray B Emeneau. India as a lingustic area. *Language*, 1956.

[EN16] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the SIGSAC Conference on Computer and Communications Security*, pages 1388–1401. ACM, 2016.

[EN20] Steven Englehardt and Arvind Narayanan. Openwpm framework. https://github.com/mozilla/OpenWPM, 2020.

[ERE+15] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web*, pages 289–299, 2015.

[Eur18] European Commission. Rules for the protection of personal data inside and outside the eu. https://ec.europa.eu/info/law/law-topic/data-protection_en/, 2018.

[EZ21] Eleonora Esposito and Sole Alba Zollo. 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*, 2021.

[FBB21] Tracie Farrell, Mehmet Bakir, and Kalina Bontcheva. Mp twitter engagement and abuse post-first covid-19 lockdown in the uk: White paper. *arXiv preprint arXiv:2103.02917*, 2021.

[FDC+18] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2018.

[FS19] Tamara Fuchs and Fabian SchÄfer. Normalizing misogyny: Hate speech and verbal abuse of female politicians on japanese twitter. In *Japan forum*, pages 1–27. Taylor & Francis, 2019.

[Fut17] British Future. 52 minority mps to sit in 'most diverse uk parliament ever'. https://www.britishfuture.org/52-minority-mps-to-sit-in-most-diverse-uk-parliament-ever/, 2017.

[FW20] Deen Freelon and Chris Wells. Disinformation as political communication, 2020.

[G+10] Hongyu Gao et al. Detecting and characterizing social spam campaigns. In *Proc. ACM SIGCOMM IMC*, 2010.

[GBG+19] Mark A Greenwood, Mehmet E Bakir, Genevieve Gorrell, Xingyi Song, Ian Roberts, and Kalina Bontcheva. Online abuse of uk mps from 2015 to 2019. *arXiv preprint arXiv:1904.11230*, 2019.

[GBHVH13] Todd Graham, Marcel Broersma, Karin Hazelhoff, and Guido Van'T Haar. Between broadcasting political messages and interacting with voters: The use of twitter during the 2010 uk general election campaign. *Information, Communication & Society*, 2013.

[GBR+20] Genevieve Gorrell, Mehmet E Bakir, Ian Roberts, Mark A Greenwood, and Kalina Bontcheva. Which politicians receive abuse? four factors illuminated in the uk general election 2019. *EPJ Data Science*, 2020.

[GC18] Peter Geoghegan and Jenna Corderoy. Revealed: Arron banks brexit campaign's secret meetings with cambridge analytica. https://www.opendemocracy.net/uk/brexitinc/peter-geoghegan-jenna-corderoy/revealed-arron-banks-brexit-campaign-had-more-meetings-w, 2018.

[GDFMGM18] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, 2018.

[GGR+18] Genevieve Gorrell, Mark A Greenwood, Ian Roberts, Diana Maynard, and Kalina Bontcheva. Twits, twats and twaddle: Trends in online abuse towards uk politicians. In *Twelfth international AAAI conference on web and social media*, 2018.

[GGZY+20] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, 2020.

[GIM+99] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, 1999.

[GJB16] Todd Graham, Dan Jackson, and Marcel Broersma. New platform, old habits? candidates use of twitter during the 2010 british and dutch general election campaigns. *New media & society*, 2016.

[GM18] Sascha Göbel and Simon Munzert. Political advertising on the wikipedia marketplace of information. In *Social Science Computer Review*. SAGE, 2018.

[GMBG10] Will J Grant, Brenda Moon, and Janie Busby Grant. Digital dialogue? australian politicians' use of the social network tool twitter. *Australian journal of political science*, 2010.

[GSAG19] Aakriti Gupta, Sunil Kumar Singh, Kabir Ahuja, and Ankit Gupta. Good morning turning to spam morning. In *ICICCT*, 2019.

[GT18] Kiran Garimella and Gareth Tyson. Whatapp doc? a first look at whatsapp public group data. In *ICWSM*, 2018.

[Guh17] Ramachandra Guha. *India after Gandhi: The history of the world's largest democracy*. Pan Macmillan, 2017.

[Gup70] Jyotirindra Das Gupta. *Language Conflict and National Development: Group Politics and National Language Policy in India*. Univ of California Press, 1970.

[GVM+21] Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. An expert annotated dataset for the detection of online misogyny. In *Proceedings of European Chapter of the Association for Computational Linguistics*, 2021.

[GW17] Venkata Rama Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2017.

[Hal12] Scott A Hale. Net increase? cross-lingual linking in the blogosphere. *Journal of Computer-Mediated Communication*, 2012.

[Hal14] Scott A Hale. Multilinguals and wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, 2014.

[Hal15] Scott A Hale. Cross-language wikipedia editing of okinawa, japan. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[Hal16] Scott A Hale. User reviews and language: how language influences ratings. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2016.

[HBKH20] Ehsan Ul Haq, Tristan Braud, Young D Kwon, and Pan Hui. A survey on computational politics. *IEEE Access*, 2020.

[HCC11] Lichan Hong, Gregorio Convertino, and Ed H Chi. Language matters in twitter: A large scale study. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[HG10] Brent Hecht and Darren Gergle. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010.

[Hin16] Gaby Hinsliff. Trash talk: How twitter is shaping the new politics. The Guardian, 2016. Available at https://bit.ly/2aFWEBu.

[HLAH18] Shiqing He, Allen Yilun Lin, Eytan Adar, and Brent Hecht. The tower of babel.jpg: Diversity of visual encyclopedic knowledge across wikipedia language editions. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[HMSG08] Lucie N Hutchins, Sean M Murphy, Priyam Singh, and Joel H Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 2008.

[Hou17] House of Commons Library. Political disengagement in the uk: Who is disengaged? Parliament research briefing, August 2017.

[HRG01] William T Harding, Anita J Reed, and Robert L Gray. Cookies and web bugs: What they are and how they work together. *Information Systems Management*, 18(3):17–24, 2001.

[HS19] Xuehui Hu and Nishanth Sastry. Characterising third party cookie usage in the eu after gdpr. In *Proceedings of the 10th ACM Conference on Web Science*, pages 137–141, 2019.

[INPG10] Takashi Iba, Keiichi Nemoto, Bernd Peters, and Peter A Gloor. Analyzing the creative editing behavior of wikipedia editors: Through dynamic social network analysis. In *Social and Behavioral Sciences*. Elsevier, 2010.

[IPS18] IPSA. The scheme of mps' business costs and expenses, 2018. Available at https://bit.ly/2UaM6jv.

[IRI10] IRIS Service. Advice for members and their staff, data protection act 1998, 2010. Available at https://bit.ly/2FJ9H0O.

[Jac08] Nigel Jackson. Representation in the blogosphere: Mps and their new constituents. *Parliamentary Affairs*, 2008.

[JCG17] Sam Joiner, Stefano Ceccon, and Louis Goddard. General election: How the middle fell out of british politics? The Times, UK, 2017.

[JL11] Nigel Jackson and Darren Lilleker. Microblogging, constituency service and impression management: Uk mps and the use of twitter. *The Journal of Legislative Studies*, 2011.

[JM09] Carter Jernigan and Behram FT Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[Jon17] Matt Jones. How whatsapp reduced spam while launching end-to-end encryption, 2017. usenix.org/conference/enigma2017/conference-program/presentation/jones.

[Jun16] Andreas Jungherr. Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, 2016.

[Kag18] Kaggle. Jigsaw toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/, 2018.

[KB15] Sunder Katwala and Steve Ballinger. The race for representation: How ethnic diversity became the 'new normal'in british politics. *London: British Future*, 2015.

[KE19] Amir Karami and Aida Elkouri. Political popularity analysis in social media. In *International conference on information*, pages 456–465. Springer, 2019.

[KHN+18] Young Mie Kim, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, Richard Heinrich, Robyn Baragwanath, and Garvesh Raskutti. The stealth media? groups and targets behind divisive issue campaigns on facebook. *Political Communication*, pages 1–27, 2018.

[Kin91] Gary King. Constituency service and incumbency advantage. *British Journal of Political Science*, 1991.

[KMBP18] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M Pujol. Whotracks. me: Monitoring the online tracking landscape at scale. *CoRR, abs/1804.08959*, 2018.

[Kra97] Jonathan S Krasno. *Challengers, competition, and reelection: Comparing Senate and House elections*. Yale University Press, 1997.

[KS18] Lucie-Aimée Kaffee and Elena Simperl. Analysis of editors' languages in wikidata. In *Proceedings of the 14th International Symposium on Open Collaboration*. ACM, 2018.

[KSPC07] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in wikipedia. In *Human Factors in Computing Systems*, 2007.

[LB12] Cristina Leston-Bandeira. Studying the relationship between parliament and citizens. *The Journal of Legislative Studies*, 2012.

[Lev19] Sam Levin. How consumers from tier ii and iii cities are powering india's growth, 2019.

[LGG18] Preethi Lahoti, Kiran Garimella, and Aristides Gionis. Joint non-negative matrix factorization for learning ideological leaning on twitter. In *Proceedings of the 11th International Conference on Web Search and Data Mining*, pages 351–359. ACM, 2018.

[LKM13] Darren G Lilleker and Karolina Koc-Michalska. Online political communication strategies: Meps, e-representation, and self-representation. *Journal of Information Technology & Politics*, 2013.

[Lok18] CSDS Lokniti. How widespread is whatsapp's usage in india? Live Mint, 2018.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[LSS+15] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22nd SIGSAC Conference on Computer and Communications Security*, pages 554–566. ACM, 2015.

[M+20] Alexandros Mittos et al. Online harms: A meta-tool for abusive speech detection. https://github.com/amittos/OnlineHarms-Metatool, 2020.

[MBGM15] Dhiraj Murthy, Sawyer Bowman, Alexander J Gross, and Marisa McGarry. Do we tweet differently from our mobile devices? a study of language differences on mobile and web-based twitter platforms. *Journal of Communication*, 2015.

[McN11] Brian McNair. *An introduction to political communication*. Routledge, 2011.

[Med20] MediaBias. Media bias/fact check. mediabiasfactcheck.com/, 2020.

[Med21] Vox Media. Coral toxic comments: Leveraging google's perspective api. https://legacy.docs.coralproject.net/talk/toxic-comments/, accessed on 20 March 2021, 2021.

[MM12] Panagiotis T Metaxas and Eni Mustafaraj. Social media and the elections. *Science*, 2012.

[Mul15] Lincoln Mullen. textreuse: Detect text reuse and document similarity. *rOpenSci*, 2015.

[MW14] Solomon Messing and Sean J Westwood. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, 2014.

[MWWD06] John McCarthy, Peter Wright, Jayne Wallace, and Andy Dearden. The experience of enchantment in human–computer interaction. *Personal and ubiquitous computing*, 10(6):369–378, 2006.

[NFKN19] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. Reuters Institute Digital News Report 2019 , 2019.

[NG04] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[NKJ+13] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of the Symposium on Security and Privacy*, pages 541–555. IEEE Computer Society, 2013.

[NLK+13] Jessica J Neff, David Laniado, Karolin E Kappler, Yana Volkovich, Pablo Aragón, and Andreas Kaltenbrunner. Jointly they edit: Examining the impact of community identification on political interaction in wikipedia. *PloS one*, 2013.

[NTC15] Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. Audience and the use of minority languages on twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[NW90] Philip Norton and David Wood. Constituency service by members of parliament: Does it contribute to a personal vote? *Parliamentary Affairs*, 1990.

[NW93] Philip Norton and David M Wood. Back from westminster: Constituency service by british members of parliament. *Lexington: The University Press of Kentucky*, 1993.

[OMDC14] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. Selling off privacy at auction. In *Network and Distributed System Security Symposium (NDSS)*, 2014.

[Oom18] Jeroen Ooms. Google's compact language detector 2, 2018.

[P+19] Fahad Pervaiz et al. An assessment of SMS fraud in pakistan. In *Proc. ACM CCS*, 2019.

[Par13] UK Parliament. Members' names information service api. https://data.parliament.uk/membersdataplatform/, accessed on 28 April 2021, 2013.

[Par16] UK Parliament. Hate crime: Abuse, hate and extremism online. https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60902.htm, 2016.

[Par17] UK Parliament. House of commons library, general election 2017: Full results and analysis. https://commonslibrary.parliament.uk/research-briefings/cbp-7979/, 2017.

[Par19] UK Parliament. Written ministerial statement by rt hon oliver dowden mp: Update on tackling intimidation in public life. https://questions-statements.parliament.uk/written-statements/detail/2019-11-05/hcws100, 2019.

[Par21] UK Parliament. Draft online safety bill. https://www.gov.uk/government/publications/draft-online-safety-bill, 2021.

[Pat19a] Keshav Patel. Indian social media politics-new era of election war. *Research Chronicler, Forthcoming*, 2019.

[Pat19b] Priya Pathak. Whatsapp is banning 2 million accounts every month. India Today, 2019.

[PBJB15] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, The University of Texas at Austin, 2015.

[PDP+17] Elias P. Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos P. Markatos. The long-standing privacy debate: Mobile websites vs mobile apps. In *Proceedings of the 26th International Conference on World Wide Web*, pages 153–162, 2017.

[Pet19] Katarina Pettersson. Freedom of speech requires actions: Exploring the discourse of politicians convicted of hate-speech against muslims. *European Journal of Social Psychology*, 2019.

[PH14] Ferran Pla and Lluís-F Hurtado. Political tendency identification in twitter using sentiment analysis techniques. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers*, pages 183–192, 2014.

[Pia20] James A Piazza. Politician hate speech and domestic terrorism. *International Interactions*, 2020.

[PKM19] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *Proceedings of the 28th International Conference on World Wide Web*, 2019.

[PKQSLCC19] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and monitoring hate speech in twitter. *Sensors*, 2019.

[PL18] Paul Hilder Paul Lewis. Leaked: Cambridge analytica's blueprint for trump victory. https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory, 2018.

[PP11] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 430–438. ACM, 2011.

[PP19] Joyojeet Pal and Anmol Panda. Twitter in the 2019 indian general elections: Trends of use across states and parties. *Economic and Political Weekly*, 54, 2019.

[PRCW20] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. Quantifying engagement with citations on wikipedia. In *World Wide Web Conference*, 2020.

[QBL+19] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[Rao19] H Naresh Rao. The role of new media in political campaigns: A case study of social media campaigning for the 2019 general elections. *Asian Journal of Multidimensional Research (AJMR)*, 2019.

[RJC+19] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. Challenges in the decentralised web: The mastodon case. In *Proceedings of the Internet Measurement Conference*, 2019.

[RMGB20] Julio CS Reis, Philipe Melo, Kiran Garimella, and Fabrício Benevenuto. Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Misinformation Review*, 2020.

[RMS+19] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. (mis) information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*, pages 818–828, 2019.

[Rt 02] Rt Hon Robin Cook MP. Speech at the yougov reviving democracy conference, qeii centre, london, April 2002.

[RvS21] Roderik Rekker and Joost van Spanje. Hate speech prosecution of politicians and its effect on support for the legal system and democracy. *British Journal of Political Science*, pages 1–22, 2021.

[S+90] David W Schumann et al. Predicting the effectiveness of different strategies of advertising variation: A test of the repetition-variation hypotheses. *J. Consumer Research.*, 1990.

[Sas12] Nishanth Ramakrishna Sastry. How to tell head from tail in user-generated content corpora. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[SB17] Isaac Stanley-Becker. The center in british politics has all but disappeared, leaving the country as polarized as the u.s. Washington Post, June 2017.

[SB20] Alexandra A Siegel and Vivienne Badaan. # no2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, 2020.

[Sco19] Jennifer Scott. Women mps say abuse forcing them from politics. bbc news. https://www.bbc.co.uk/news/election-2019-50246969, 2019.

[SIIK19] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. Talon: An automated framework for cross-device tracking detection. In *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, pages 227–241, 2019.

[SKV10] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *CCS*, 2010.

[SLTVSV17a] Criag Silverman, Jane Lytvynenko, Lam Thuy Vo, and Jeremy Singer-Vine. Data, analytic code, and findings related to the buzzfeed news article, 2017. https://github.com/BuzzFeedNews/2017-08-partisan-sites-and-facebook-pages.

[SLTVSV17b] Criag Silverman, Jane Lytvynenko, Lam Thuy Vo, and Jeremy Singer-Vine. Inside the partisan fight for your news feed, 2017. https://www.buzzfeednews.com/article/craigsilverman/inside-the-partisan-fight-for-your-ne

[Spe15] Speaker's Commission. Open up! report of the speaker's commission on digital democracy, Jan 2015. Available at https://bit.ly/1HMrm33.

[SS18] Patrick Svitek and Haley Samsel. Ted cruz says cambridge analytica told his presidential campaign its data use was legal. https://www.texastribune.org/2018/03/20/ted-cruz-campaign-cambridge-analytica/, 2018.

[Sta08] James Stanyer. Elected representatives, online self-presentation and the personal vote: Party, personality and webstyles in the united states and united kingdom. *Information, Community & Society*, 2008.

[Sta17] Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *AAAI ICWSM*, 2017.

[STDE19] Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. The wisdom of polarized crowds. In *Nature human behaviour*. Nature Publishing Group, 2019.

[TESU15] Gareth Tyson, Yehia Elkhatib, Nishanth Sastry, and Steve Uhlig. Are people really social in porn 2.0? In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[The00] The Federal Trade Commission. Ftc sues failed website, toysmart.com, for deceptively offering for sale personal information of website visitors. https://www.ftc.gov/news-events/press-releases/2000/07/ftc-sues-failed-website-toysmartcom-deceptively-offering-sale, 2000.

[TP10] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[TR20] Thread-Reader. Fake flipkart website, 2020. bit.ly/fake-FK.

[TS12] Neil Thurman and Steve Schifferes. The future of personalization at news websites: Lessons from a longitudinal study. *Journalism Studies*, 2012.

[VAA+21] Yash Vekaria, Vibhor Agarwal, Pushkal Agarwal, Sangeeta Mahapatra, Sakthi Balan Muthiah, Nishanth Sastry, and Nicolas Kourtellis. Differential tracking across topical webpages of indian news media. In *Proceedings of the ACM WebSci*, 2021.

[Vac13] Cristian Vaccari. *Digital politics in Western democracies: A comparative study*. JHU Press, 2013.

[vARKL18] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, 2018.

[VBM21] Bertie Vidgen, Emily Burden, and Helen Margetts. Understanding online hate: Vsp regulation and the broader context. http://bit.ly/ofcom-turing2021, 2021.

[VFD+17] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, 2017.

[VSDV15] Joost Van Spanje and Claes De Vreese. The good, the bad and the voter: The impact of hate speech prosecution of a politician on electoral support for his party. *Party Politics*, 2015.

[wbCG19] whotracks.me by Cliqz and Ghostery. Bringing transparency to online tracking., 2019. github.com/cliqz-oss/whotracks.me/tree/master/whotracksme/data/assets.

[WBJ+20] Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 2020.

[WH16] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 2016.

[Whe17] Jonathan Wheatley. The polarisation of party supporters since 2015 and the problem of the 'empty centre' – in maps. London School of Economics Blog, June 2017.

[Wil19] John Wilander. Intelligent tracking prevention 2.3. https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/, 2019.

[WM20] Stephen Ward and Liam McLoughlin. Turds, traitors and tossers: the abuse of uk mps via twitter. *The Journal of Legislative Studies*, 2020.

[Wor10] World Wide Web Consortium (W3C). Same origin policy, 2010.

[WTD17] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, 2017.

[WTSC16] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering*, 28(8):2158–2172, 2016.

[YB10] Sarita Yardi and Danah Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, 2010.

[YRS+10] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 2010.

[Zau10] Christoph Zauner. Implementation and bench-marking of perceptual image hash functions. *pHash.org*, 2010.

[ZCB+18] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018.

[ZCDC+19] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation

warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 2019.

[ZCS+19] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science*, 2019.

[ZMN+19] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, 2019.

[ZWB+19] Ark Fangzhou Zhang, Ruihan Wang, Eric Blohm, Ceren Budak, Lionel P Robert Jr, and Daniel M Romero. Participation of new editors after times of shock on wikipedia. In *AAAI ICWSM*, 2019.

[ZYCG07] Shenghuo Zhu, Kai Yu, Yun Chi, and Yihong Gong. Combining content and link for classification using matrix factorization. In *Proceedings of the 30th SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2007.