Understanding Wikipedia Practices Through Hindi, Urdu, and English Takes on an Evolving Regional Conflict

MOLLY G. HICKMAN, Computer Science, Virginia Tech, USA
VIRAL PASAD, Computer Science, Virginia Tech, USA
HARSH KAMALESH SANGHAVI, Industrial and Systems Engineering, Virginia Tech, USA
JACOB THEBAULT-SPIEKER, The Information School, University of Wisconsin - Madison, USA
SANG WON LEE, Computer Science, Virginia Tech, USA

Wikipedia is the product of thousands of editors working collaboratively to provide free and up-to-date encyclopedic information to the project's users. This article asks to what degree Wikipedia articles in three languages — Hindi, Urdu, and English — achieve Wikipedia's mission of making neutrally-presented, reliable information on a polarizing, controversial topic available to people around the globe. We chose the topic of the recent revocation of Article 370 of the Constitution of India, which, along with other recent events in and concerning the region of Jammu and Kashmir, has drawn attention to related articles on Wikipedia. This work focuses on the English Wikipedia, being the preeminent language edition of the project, as well as the Hindi and Urdu editions. Hindi and Urdu are the two standardized varieties of Hindustani, a lingua franca of Jammu and Kashmir. We analyzed page view and revision data for three Wikipedia articles to gauge popularity of the pages in our corpus, and responsiveness of editors to breaking news events and problematic edits. Additionally, we interviewed editors from all three language editions to learn about differences in editing processes and motivations, and we compared the text of the articles across languages as they appeared shortly after the revocation of Article 370. Across languages, we saw discrepancies in article tone, organization, and the information presented, as well as differences in how editors collaborate and communicate with one another. Nevertheless, in Hindi and Urdu, as well as English, editors predominantly try to adhere to the principle of neutral point of view (NPOV), and for the most part, the editors quash attempts by other editors to push political agendas.

CCS Concepts: • **Human-centered computing** → **Wikis**.

Additional Key Words and Phrases: Wikipedia; collaboration; Jammu & Kashmir

1 INTRODUCTION

Wikipedia defines itself as a multilingual online encyclopedia, created and maintained as an open collaboration project using a wiki-based editing system. One of its core principles is to maintain consensus around a neutral point of view in Wikipedia articles [23, 51]. The Wikimedia Foundation and volunteer editor community have fostered over 250 language editions of Wikipedia, the most frequently edited and visited of them being the English edition [21]. When information is not shared between different language editions of Wikipedia, it prevents access to a larger variety of content for monolingual users [23], creating a *knowledge gap* between language editions. To counteract this, Wikipedia offers tools and guidance for translating or requesting translations into English from another language [49] and vice versa [48].

This knowledge gap problem may be exacerbated in the case of articles related to polarizing topics. Information seekers who are in proximity to a conflict, and who only access media coverage of said conflict in their local languages, may get an incomplete or biased picture of events [22]. Adversarial actors have been known to exploit Wikipedia pages and to manipulate search engine

Authors' addresses: Molly G. Hickman, mollygh@vt.edu, Computer Science, Virginia Tech, USA, Blacksburg, Virginia; Viral Pasad, viralpasad@vt.edu, Computer Science, Virginia Tech, USA, Blacksburg, Virginia; Harsh Kamalesh Sanghavi, harshks@vt.edu, Industrial and Systems Engineering, Virginia Tech, USA, Blacksburg, Virginia; Jacob Thebault-Spieker, jacob.thebaultspieker@wisc.edu, The Information School, University of Wisconsin - Madison, USA, Madison, Wisconsin; Sang Won Lee, sangwonlee@vt.edu, Computer Science, Virginia Tech, USA, Blacksburg, Virginia.

Fig. 1. A map of the Jammu & Kashmir region. Image credit: BBC [31]

results to present a biased, false, or purposefully misleading version of events, thereby disseminating their propaganda to the broader internet [19, 20]. Therefore, broadly, this paper focuses on how Wikipedia editors cope with the challenges in sustaining a neutral point of view (NPOV).

The goal of this paper is to understand how the editors of two still-growing, smaller language editions of Wikipedia, Hindi and Urdu, create a collective memory around the long-standing and evolving conflict in Jammu and Kashmir, compared to one another and to the editors of the corresponding articles in English. A confluence of factors means this content risks violating the NPOV policy and reflecting skewed narratives about the conflict in Jammu and Kashmir. Specifically, because this is a politically contentious issue, articles on the conflict and related topics in various language editions risk being targeted by partisan or biased groups looking to advance an agenda. Further exacerbating this risk is the relative size of smaller language editions like Hindi and Urdu, which may not have the resources to counteract coordinated efforts to violate the NPOV policy. Thus, to understand how these smaller language editions represent this controversial topic, our analysis focuses on three dimensions: (a) how well each language edition kept up with current events, (b) how popular the articles in our corpus were over time, and (c) to what extent each was confronted with potentially adversarial edits. To these ends, we conducted time series analyses of page view and revision data, and compared the content and composition of each article across languages. In addition, we interviewed editors to learn how their writing practices vary from what we know from previous works, and to learn how they maintain a neutral point of view (NPOV) on a topic that is prone to bias. We found many ways in which the Hindi and Urdu editors that we interviewed differed from their English counterparts in the way they collaborate, as well as one key similarity: all of the editors who were willing to participate in interviews, from all three language editions, sought to adhere to NPOV policies when contributing to articles related to the conflict.

1.1 Jammu and Kashmir Conflicts

Jammu and Kashmir (J&K) is a union territory administered and claimed by India. Ever since gaining independence from the British, Pakistan and India have been in conflict over Kashmir, a disputed region in the northern part of both countries, where they border each other. While both countries only control a part of the former princely state, both claim Jammu and Kashmir in

its entirety [25], and most Kashmiris want complete freedom for Kashmir as a separate country [24, 39]. Three wars and several other conflicts have yielded the current line of control, which demarcates the regions administered by each of three nations: India, Pakistan, and China. An insurgency began to proliferate in India-administered Kashmir in the late 1980s. Three-quarters of the region's population adheres to Islam, which is the state religion of Pakistan. Thanks to Article 370 of the Constitution of India, J&K was granted autonomy in 1954. Therefore, all laws passed by the Indian Parliament required the assent of the J&K government for them to apply. However, on August 5, 2019, the President of India issued an order overriding a previous presidential order and nullifying all the provisions of autonomy granted to the state (the Revocation of Article 370)[1]. Thereafter, the Home Minister of India introduced a reorganization bill in the Indian Parliament to divide the state into two union territories [34]. This situation has brought unrest into the region and further destabilized the already tenuous relations between India and Pakistan. More on the conflict can be found in the appendix. Recent events (namely, the Pulwama Attack — see appendix A — and the Revocation of Article 370) and their rippling effects motivated us to study the ongoing conflict in Kashmir and how it is represented on Wikipedia. We chose to study Wikipedia articles written in three primary languages spoken in the region of the conflict: Hindi, Urdu, and English. Hindi is the most widely spoken language in India, and Pakistan's national language is Urdu; linguistically, both are actually standardized varieties of Hindustani, a lingua franca of J&K. English is widely used in both countries, where it enjoys official status alongside other languages [15, 16].

1.2 Neutral Point of View

The way that different language editions represent such a controversial topic creates a difficult definitional problem at the core of our research here: what is the *right* way to fairly represent this controversy? After all, points of view (POVs) may differ on how events transpired, and different facets of the event may be understood differently by different groups. Therefore, for the purposes of this paper, we adopt Wikipedia's operational guidelines on a "neutral point of view" (NPOV), which boil down to the following statement: *"Articles must not take sides, but should explain the sides, fairly and without editorial bias. This applies to both what you say and how you say it* [14]." NPOV is one of Wikipedia's three core content policies, along with verifiability[2] and no original research[3]. The NPOV policy is composed of nine guidelines. We provide an introduction to the three most salient guidelines to our work — which we refer to throughout the paper — here:

- *Due and undue weight*: This guidance refers to the requirement that articles treat all relevant viewpoints "in proportion to the prominence of each viewpoint in the published, reliable sources [14]."
- *Impartial tone*: "Wikipedia aims to describe disputes, but not engage in them [14]." Editors must not give the impression that they endorse nor reject any particular POV.
- *Words to watch*: Editors are asked to "eliminate expressions that are flattering, disparaging, vague, or clichéd, or that endorse a particular point of view (unless those expressions are part of a quote from a noteworthy source) [14]."

We also reviewed the corresponding NPOV page in the Hindi and Urdu language editions as they appeared on May 29, 2020. The Hindi page on NPOV includes the first two guidelines, but

[1]The third clause of Article 370 states: "Notwithstanding anything in the foregoing provisions of this article, the President may, by public notification, declare that this article shall cease to be operative or shall be operative only with such exceptions and modifications and from such date as he may specify," provided that the Constituent Assembly of the State has made its recommendation as specified in clause (2) (https://en.wikipedia.org/wiki/Article_370_of_the_Constitution_of_India#Original_text).

[2]"Material challenged or likely to be challenged, and all quotations, must be attributed to a reliable, published source [13]."

[3]"Wikipedia does not publish original thought: all material in Wikipedia must be attributable to a reliable, published source [13]."

not *Words to watch*. The Urdu page on NPOV is very short and includes none of the itemized guidelines present in the English and Hindi versions — only the basic definition of NPOV and how it works together with the other two core content policies. We chose to use the English Wikipedia's definition of NPOV as our operational definition in this work, since it is the most comprehensive. The English Wikipedia is the largest, and as seen in our interviews, something of a de facto standard.

1.3 Research Questions and Contributions

Editor behavior and adherence to Wikipedia norms are central to how the J&K conflict is represented on Wikipedia, raising the question: *How do editors' Wikipedia practices in the three languages vary in general, and how do editors uphold the NPOV principle?* We approached this question from the following perspectives.

- How do the Wikipedia communities and editors' practices vary across the Hindi, Urdu, and English Wikipedias?
- To what extent has a neutral point of view (NPOV) been sought on a topic that invites edits with an agenda?

While the English language edition of Wikipedia has been widely studied, editor behavior, motivation, and difficulties in non-English/non-Western language editions, like those we study here, are less deeply understood. By contrasting the Hindi and Urdu language editions with one another, as well as with the English language edition, and specifically focusing on a polarizing topic, our work develops insights into all three communities of editors — their motivations, their collaborative practices, and their struggles with potentially adversarial editors. We further contribute to the body of research on how people create information and combat misinformation online, especially in response to events that draw many readers to Wikipedia, seeking a trusted and fair account of an event. We also hope that our quantitative methods may help other researchers seeking to compare the scale, responsiveness, and composition of Wikipedia language editions. In summary, our contributions are:

- We characterize the scale difference in Wikipedia communities in English and two South Asian languages.
- We extend previous work understanding Wikipedia as a source of timely information.
- We uncover evidence of challenges in sustaining a neutral point of view (NPOV) on articles related to controversial topics (like regional conflicts).
- We show that, while the goals and values of editors are similar across language editions, their sources and collaborative practices differ.

2 RELATED WORKS

Wikipedia has been an active research topic and helped us understand the participatory nature of collaborative knowledge sharing enabled in Web 2.0 on a global scale. Our study builds on previous research into the multilingual nature of Wikipedia, cultural differences between editors, how editors respond to breaking news events, and what motivates people to edit Wikipedia articles, as well as disinformation and the potential of missing or misleading information to influence public opinion.

A number of researchers have studied multilingual aspects of Wikipedia. Hecht et al. [23] have explained how Wikipedia is a global source of world knowledge and how knowledge diversity can be exploited to create "culturally aware" and "hyper-lingual" applications. Their other work [22] highlights how self-focus in community-maintained knowledge repositories can be explained.

Hale [21] has examined the roles of users editing multiple language editions of Wikipedia articles. He discusses how multilingual editors bring over more context, sources, and perspective from their main language articles to articles in other languages. Thus, they reduce the level of self-bias on Wikipedia. They also show that multilingual editors often edit the same article in their primary and non-primary languages.

Along similar lines, Pfeil et al. [37] have studied cultural differences between Wikipedia editors (specifically on the French, German, Japanese, and Dutch Wikipedias). They found several factors that differed between the cultures they studied, including reluctance to delete information, likelihood to collaborate with other editors, and propensity for spelling mistakes.

A 2017 study of multilingual students' Wikipedia habits by Soler-Adillon and Freixa showed that their subjects favor the English Wikipedia when looking for information, even if English is not the language in which they are most fluent [41]. Their subjects were able to read and write in at least Catalan, Spanish, and English. Our work here suggests these findings extend to Hindi and Urdu speakers who are also proficient in English.

Many more have focused exclusively on the English Wikipedia. Bryant et al. [4] studied how Wikipedia users' practices change as they become more experienced. At first, they tend to operate on the periphery, making marginal edits on topics they know about. As a user gains more experience, Wikipedia as a whole becomes more important than any individual article: users then tend to adopt specialized roles (e.g., administrator) and use the facilities for discussing with other editors and tracking changes to articles.

Keegan [27] has studied how Wikipedia editors collectively respond to breaking news events, and Keegan et al. [46] studied the same phenomenon as it related specifically to the Black Lives Matter movement. New events, even if not directly related to the article at hand, will spur more page views, as well as edits that put the topic in the context of the new event. With every tragic new death of a black person at the hands of police, the *Black Lives Matter* page will see a spike in attention and edits, and often, other articles in the web of articles related to the Black Lives Matter movement will also be viewed and edited.

Relatedly, Xie et al. [6] studied the challenges of analyzing the impact of any given external event on Wikipedia page view data; essentially, that seasonal trends and unrelated events make for a lot of noise. Conversely, we leverage these spikes due to external events as proxies for external events, a view supported by Xie et al. and others. For instance, Moyer et al. [29] show that Wikipedia page views can even predict passive user activity on non-Wikipedia sites like Reddit.

Farič and Potts have presented a study on the behavior and motivations of contributors to health-related articles on Wikipedia [18]. One of the key findings from their interviews is that contributors do not need to be experts on a given topic; lay people can also contribute to articles on Wikipedia, which is a theme that also echoes through our findings.

Others have studied how disinformation propagates on platforms like Wikipedia and the broader internet. Flores-Saviaga and Savage [19] have studied computational propaganda against Latinx communities in the United States. They found that a majority of political content about Latinx people on Reddit was created by extremists and trolls pushing dis- or misinformation. Flores-Saviaga and Savage referred to the lack of "neutral" voices on this topic as a "data void," a term coined by Golebiewski and boyd [20] to describe search terms/topics that return "limited, non-existent, or deeply problematic" results.

Of particular relevance to our study is one kind of data void identified by Golebiewski and boyd: the kind that can be weaponized following a breaking news event. Adversarial actors create content before trustworthy sources so that search results present a biased, false, or purposefully misleading version of events. Adversarial actors are known to exploit search engines so that the information

available in the first few results (often Wikipedia pages) is misleading or sows discord — particularly in the aftermath of breaking news events [19, 20].

Starbird et al. offer an understanding of disinformation as something that human crowds do, not just something that is done to human crowds by "bots" or nebulous government entities [42]. Sometimes, disinformation is about what is *not* being said; i.e., not telling the whole story, as opposed to outright lies. Wikipedians are a human crowd, and crowds of bad actors certainly can use Wikipedia to spread disinformation (sometimes just by flooding articles with weird edits or deletions to muddy the water or waste editors' time, as an interviewee told us). However, coordinated attempts to spread disinformation are often nipped in the bud thanks to good-faith editors' diligence.

3 METHOD

Here, we focus on comparing Wikipedia editors' behavior, motivations, and reactions to significant events related to the polarizing conflict in J&K between the Hindi, Urdu and English Wikipedias. First, to understand the scale difference among the three languages, we conducted quantitative analysis on each article and compared various metrics, including readership, article length, and responsiveness (both to events and to problematic edits). Second, to investigate how NPOV is sought in the target articles and editors' practices, we took a qualitative approach to examining the Wikipedia articles and understanding the practices of editors who contributed to the target articles. Our work examines how editors in different language editions cover this controversial topic. In particular, it is possible that different language editions produce articles about this controversial topic that violate Wikipedia's NPOV policy. This may occur for a variety of reasons, including the relatively small number of editors and focused efforts to represent a controversial topic in line with a particular agenda or worldview. We therefore refer to these as *vNPOV ("violating NPOV")* edits or practices throughout the paper, reflecting the salience of the NPOV policy in the Wikipedia community.

In the remainder of this section, we describe our method in detail. Our code and the text of the articles discussed herein (originals and translations) can be found at https://github.com/ViralNotPrasad/cscw-wiki.

3.1 Target Wiki Pages

Based on our topic of interest — the conflicts in Jammu and Kashmir — we chose to study the following Wikipedia pages that are closely related to the topic (page creation dates in parentheses):

- (**A1**) *Article 370 of the Constitution of India*
  - **A1-EN** English (2006-02-13): https://en.wikipedia.org/wiki/Article_370_of_the_Constitution_of_India
  - **A1-HI** Hindi (2011-03-12) [4]: https://hi.wikipedia.org/wiki/ अनुच्छेद _३७०
  - **A1-UR** Urdu (2018-09-27): https://ur.wikipedia.org/wiki/370_ آئین_ہند_کی_دفعہ
- (**A2**) *Pulwama attack*
  - **A2-EN** English (2019-02-15): https://en.wikipedia.org/wiki/2019_Pulwama_attack
  - **A2-HI** Hindi (2019-02-15): https://hi.wikipedia.org/wiki/ २०१९ _पुलवामा _हमला
  - **A2-UR** Urdu (2019-02-15): https://ur.wikipedia.org/wiki/ پلوامہ_حملہ،_2019ء
- (**A3**) *Insurgency in Jammu and Kashmir*
  - **A3-EN** English (2005-05-20): https://en.wikipedia.org/wiki/Insurgency_in_Jammu_and_Kashmir
  - **A3-HI** Hindi (2011-04-14): https://hi.wikipedia.org/wiki/ जम्मू_ और _ कश्मीर _ में _ विद्रोह
- (**A4**) *Kashmir conflict*

[4] You can navigate to the Hindi and Urdu pages via their English counterparts, using the Languages menu on the left.

  - **A4-EN** English (2005-10-29): https://en.wikipedia.org/wiki/Kashmir_conflict
  - **A4-HI** Hindi (2014-01-31): https://hi.wikipedia.org/wiki/ विवाद _कश्मीर
  - **A4-UR** Urdu (2013-10-18): https://ur.wikipedia.org/wiki/ مسئلہ_کشمیر
- (**A5**) *Jammu and Kashmir Reorganization Act, 2019* (English, Hindi, Urdu)
  - **A5-EN** English (2019-08-06): https://en.wikipedia.org/wiki/Jammu_and_Kashmir_Reorganisation_Act,_2019
  - **A5-HI** Hindi (2019-08-07): https://hi.wikipedia.org/wiki/ जम्मू _और _कश्मीर _पुनर्गठन _अधिनियम ,_2019
  - **A5-UR** Urdu (2019-08-07): https://ur.wikipedia.org/wiki/ ء2019_, ایکٹ_ نو_ تنظیم_ کشمیر_ و_ جموں

Of these five articles, only *Insurgency in Jammu and Kashmir*(A3) does not have a version in all three languages — the Urdu Wikipedia lacks its own version of this article.

We gathered data on revisions and page views for each article via the MediaWiki API[5]. For page view data, we used the Python library `pageviewapi`[6], and for reversions, we relied on the `mwreverts` library[7]. The data, collected on May 21, 2020, include the editor's username or IP address, timestamp of the revision, size in bytes of the edit (negative for edits that shortened the articles), content of the edit, and a comment explaining the revision (if provided). In all Hindi and Urdu versions of articles, our dataset of edits spans the period of time from when the article was created until May 21, 2020. This is also true for the English version of the *Reorganisation Act (A5-EN)* article. Due to large differences in scale, we limit most articles in the English language edition to the 500 most recent edits. We also sampled the content of each article as it appeared at midnight on September 1, 2019, AOE, and we had the Hindi and Urdu articles translated by translators hired from Upwork to compare their contents [47].

3.2 Page Views

We examined a page view time series for each article to look at the scale of readership of each language and article, validate our choice of articles, and support later analysis of the correlation between the number of page views and the number of edits to a given article at a given time. Information seekers visit the Wikipedia page on a topic following breaking news events that relate to the topic [27, 46]. Therefore, we use page views as a proxy for the significance of an event.

3.3 Revision Data and Anonymous Edits

To answer the question of how timely editors' responses are to significant events, we computed the Pearson correlation for each article between daily page views and daily edits. These correlations let us compare how quickly articles in different languages "catch up" to current events — a strong correlation suggests that the volume of edits spikes on the same days as does the volume of page views on a given page, and that when there is little interest by readers in a page, there is also little editing. A weak correlation could mean that editors are unresponsive, but it could also mean that an article is not closely related to current events — people may visit a page for context about a given event, but the event may not require editing on the page.

In order to get a sense of the scale of potentially problematic edits in each edition, we looked at the proportion of edits by anonymous editors, as well as the number of edits that were reverted. Anonymous edits are not necessarily bad edits [45]; however, three interviewees, one from each language edition, cited IP edits as tending to push a particular POV or otherwise be problematic. Likewise, edits are not always reverted because they are bad — in fact, reversion can itself be

[5]https://www.mediawiki.org/wiki/API:Revisions
[6]https://pypi.org/project/pageviewapi/
[7]https://github.com/mediawiki-utilities/python-mwreverts

vandalism — but we saw that reversions were typically made by fast-acting regular editors in response to misguided edits.

3.4 Article Content Analysis

We analyzed articles from each of the three Wikipedia language editions (English, Hindi, and Urdu) to investigate differences in structure and content, and to check for presence of biased information. First, to easily compare the articles (in terms of information architecture, content, headings, and subheadings), we translated the Hindi and Urdu articles into English. We searched for differences in structure, such as differing heading names, as well as headings included in some articles but missing from others, to see if these adhere to the "due and undue weight" guideline. Then, in each article across all three languages, we annotated parts of the articles that violated the NPOV guidelines introduced in § 1.2, labeling those parts *vNPOV*. Then, for each *vNPOV* occurrence in a given article, we checked whether the same subject was covered in the corresponding article(s) in the other languages; if it was, we took note of how the subject was treated differently across languages. Two authors then coded each incident and extracted common themes in an iterative fashion, ultimately reaching a consensus.

3.5 Editor Pool Analysis and Interviews

We recruited Wikipedia editors from the Hindi, Urdu, and English language editions to participate in interviews. To extract unique editors from Wikipedia, we used the `XTools` web tool [52]. We identified editors of interest for the articles in our corpus on the basis of total number of edits (the more, the better) and the date of their last edit (the more recent, the better). We sent each editor a recruitment email using Wikipedia's `EmailUser` facility [8]. Editors were asked to participate in a one-hour, semi-structured interview [2]. By nature, we were unable to request interviews with editors that edited anonymously, leaving only an IP address instead of a username.

Sl No.	Age	Country	Articles	Language Proficiency	Occupation	Role	No. of Edits	Editor Since
H1(IN)	22	IN	A4-HI	Marwari, HI, EN	Media	Editor	40–50K	2015
H2(IN)	23	IN	A1-HI, A4-HI	HI, EN	Student	Editor	<5K	2019
U1(PK)	40	PK	A1-UR, A4-UR	UR, EN	Online news editor	Admin	20–30K	2013
E1(US)	37	US	A1-EN, A2-EN, A4-EN	EN, Spanish	Technical writer	Admin	2,000,000+	2004
E2(IN)	19	IN	A1-EN, A2-EN	EN, HI, Bengali	Student	Admin	10–20K	2010
E3(SB)	35	SB	A1-EN, A3-EN	Serbian, EN	Medical Scientist	Editor	50–60K	2010

Table 1. Interviewees' demographics (HI = Hindi, UR = Urdu, EN = English).

We interviewed a total of six editors for this study. Participants represented varied backgrounds; the interviewee demographics are presented in Table 1. The purpose of the interview was to understand motivations, collaborative structures, sources of conflict, potential bias on the topic, and sources of references, as well as the editing processes of these editors in each language edition of Wikipedia.

After the interview was completed, we approached the analyses informed by grounded theory-style approaches. We extracted key concepts and ideas discussed in the interviews by performing iterative rounds of thematic analysis [3]. First, we performed open coding, wherein we assigned labels or short phrases to serve as codes to summarize relevant content we read in the interview transcripts. We analyzed six hours worth of interview audio, transcribed into 1,070 messages from our six interviewees. Each message sentence was assigned a phrase that reflected the meaning of the communication. We then re-analyzed the interview transcripts along with these phrases to extract themes that could aid in conceptual understanding. Two members of the research team met regularly to compare their analysis of the themes in an iterative fashion, going back and forth until there were no disagreements. This analytic process led us to focus on the larger themes, such as

the motivations, editing processes, communications and collaborations of the interviewees, along with their adherence to the principle of NPOV.

We evaluated to what extent editors edit Wikipedia articles across different languages; which language pairs had the biggest overlap between editor pools; and whether any articles stuck out as "hubs," where many editors worked on a particular article in addition to others within the same language. We analyzed the editorship overlap in two ways.

First, we computed Jaccard coefficients [26], used previously by Twyman et al. [46] to examine the similarities between sets of editors of Wikipedia articles related to the Black Lives Matter movement. This metric describes the portion of common editors in the union of two editor pools (e.g., the set of all English articles versus the set of all Hindi articles). It is the size of the intersection of two sets (of editors, here) divided by the size of their union. The Jaccard coefficient ranges from 0 to 1, with 1 being the most similar (all in common) and 0 being the most diverse (none in common).

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Here, A and B are sets of editors who contributed to any of the Wikipedia target articles in one of the languages.

Secondly, to see if editors worked on multiple articles within one language, we calculated the percentage of each article's editor pool that also worked on one or more other articles in the same language.

4 UNDERSTANDING THE SCALE DIFFERENCE BETWEEN LANGUAGES

4.1 Consumption and Composition of the English, Urdu and Hindi Editions of Wikipedia

In order to understand the scale difference among Wikipedia articles in three different languages, we analyzed one of the target pages (A1) with various metrics. We selected A1 because it was the most viewed article in 2019 among the target pages with a significant event (the revocation of Article 370) on August 5, 2019. The order of magnitude difference among languages was substantial. (See Fig 2-(a).) This difference was still large when the accumulated view counts were normalized by the number of speakers. (See Fig 2-(b).) We observed a similar scale difference when there was a peak in daily views on the day that Article 370 was revoked (August 5, 2019). (See Fig 2-(c).) The page views on *Article 370* (A1) in English, Hindi, and Urdu peaked at 2.53 million, 380,000, and a mere 382 views, respectively, on the day Article 370 was revoked. Given that Urdu and Hindi speakers are often also proficient or fluent in English, they may contribute to the the English Wikipedia page views, as found in a previous work [41]. While the difference can be attributed to many different factors (e.g., internet access, literacy, multilingualism), it is apparent that the Wikipedia article in English is the most actively consumed, followed by Hindi, and with Urdu being much smaller than the other two language editions, even with the proximity of the topic to the two South Asian language-speaking populations.

Language	(a) Page Views[8]	(b) Normalized Page Views[9]	(c) Peak Daily View	(d) Page Length (bytes)	(e) Talk Page (bytes)[10]	(f) Number of Sources[11]	(g) Number of Contributors	(h) IP Edits (%)
English	6.4 million	1.7E-2	2.53 million	67,331	69,696	120	183	18.4
Hindi	1.1 million	3.2E-3	380 K	25,946	938	11	44	24.4
Urdu	5.9 K	1.0E-4	382	5,420	13,471	2	12	8.1

Table 2. Scale difference of *Article 370* web pages (A1), as of May 29, 2020, unless otherwise stated

For nearly every article in our corpus (A1-5), the length of the articles (measured in bytes) was the greatest in the English edition, followed by the Hindi and then the Urdu editions. This pattern was also seen in the number of contributors for each language edition. Interestingly, there was one exception in terms of article length: A5-UR (21,616 bytes) was longer than A5-HI (6,933 bytes).

This does not mirror the difference in the total number of articles in each edition (a measure of a Wikipedia's maturity): 139,246 in Hindi, 154,057 in Urdu, and 6,087,693 in English, as of this writing (May 29, 2020). The Urdu edition yielded 10% more articles than the Hindi edition. The Hindi language edition was created before the Urdu language edition, in July 2003 and January 2004, respectively. Therefore, the difference in volume is not owed to one having had more time to develop. By contrast, there were 28 unique contributors, excluding the anonymous IP editors, to the Urdu articles in our corpus and 137 contributors to the Hindi articles (A1–A5) — see Figure 2-(g) for A1. Even taking into account that there was one fewer article in Urdu in our corpus than in Hindi, it seems apparent that more people participate in the Hindi edition. Meanwhile, there were 658 unique contributors to the English articles (A1–A5), which shows the order of magnitude difference with the other two languages.

Article talk pages are meant for discussing controversial edits. They ebb and flow in size as issues are raised and resolved; that is, if editors use them. There were several articles in our corpus (A4-UR, A2-UR, A5-HI, A5-UR) where the talk page was empty for the entirety of 2019, either because the page was never initiated (no one had ever used it in the history of the article) or because all pre-2019 issues had been removed and no new issues were brought up in 2019. Those Hindi and Urdu talk pages in our corpus that did have content in 2019 had only one contributor each, with one exception, A1-HI, which had six contributors. The A1-UR talk page got relatively long in 2019 (13 kilobytes), but it hardly seems like the one contributor posted there to start a discussion; rather, it was an itemized version of the editor's interpretation of events in J&K, including "India wants to repeat the story of Palestine in Kashmir" and referring to 38% of Kashmir as administered by Pakistan, "thank God." By contrast, all the English articles had accompanying talk pages with sizes in the thousands, usually tens of thousands, of bytes throughout 2019, with many contributors.

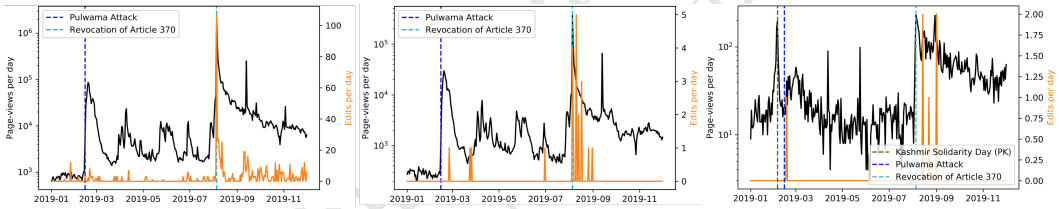

Fig. 2. Page views (black, left axis) or edits (orange, right axis) per day in 2019 for *Article 370 of the Constitution of India (A1)*, English (left), Hindi (middle), and Urdu (right). Relevant events are also displayed as dotted vertical lines: the Pulwama Attack on February 14, 2019; the revocation of Article 370 on August 5, 2019; and one more relevant event just for Urdu, Kashmir Solidarity Day, a national holiday in Pakistan. Note that the scale of the vertical axis is different for each graph.

[8](a) cumulative views between August 5 and December 31, 2019, inclusive

[7](b) divided by approximate number of native speakers of the language [10]

[10](e) maximum length in 2019. When issues are resolved, they are sometimes removed from the talk page. We recorded the maximum length of each page as a proxy for the maximum amount of conflict there was on the article in 2019, though as we will see later in the paper, not all editors use talk pages to record and resolve conflicts.

[11](f) citations and bibliography/external link items

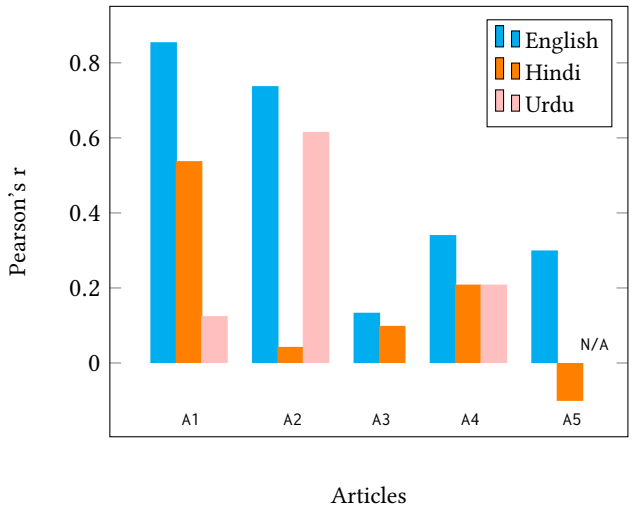

Fig. 3. Correlation between daily page views and edit time series for each article. The correlation is N/A for A5-UR because the only edits to the page occurred before there were any page views — the series did not overlap, so no correlation could be computed.

4.2 Comparison via Responsiveness of Wikipedia Pages to Relevant Events

When page views surge on an article, we tend to find that some related event has spurred readers to visit the page — and sometimes, we observe a surge only in a subset of language editions. We examined the daily views of *Article 370* (A1) over time and compared them to the three events relevant to the topic. There were local maxima in page views per day on February 14, 2019, and August 5, 2019 (the dates of the Pulwama Attack and the revocation of Article 370, respectively) for most of the articles in our corpus that existed on those dates. The only two exceptions were two Urdu articles, *Article 370* (A1-UR) and *Kashmir Conflict* (A4-UR); there was a spike on February 5, 2019 on A1-UR, instead. This can be attributed to Kashmir Solidarity Day, a national holiday in Pakistan, where Urdu is the national language [7]. Both A1-UR and A4-UR did have a smaller swell in page views a few days after the Pulwama Attack. The connection between page views and relevant events is presented in Figure 2. The fact that people sought out information on these three dates confirms that the articles in our corpus are of particular relevance to current events in J&K.

We also looked at the composition side of the wiki pages, analyzing the editing activity over time (see the orange line in Figure 2). Immediately following the revocation of Article 370 on August 5, 2019, the page on *Article 370 of the Constitution of India* in all three language editions experienced a flurry of activity involving many edits, mostly additions, explaining the action taken by the Indian government. The page on the English Wikipedia, having hovered around 37 kilobytes in length for most of 2019, rose to over 60 kilobytes (a 62.1% increase) within days of the revocation. Meanwhile, the Hindi edition went from just under 18 kilobytes to almost 26 kilobytes (a 44.4% increase), and the Urdu edition saw no significant change in article volume, staying at around 5 kilobytes. Figure 2 shows page views per day and edits per day on the three language editions of the page. Articles that are tangentially related to a breaking news event likewise experience surges in editing activity, but to a lesser degree.

In order to compare the responsiveness among English, Hindi, and Urdu editors, we examined how closely aligned the page views were with edit volumes. As noted in § 3.2, page view count

can be a good proxy for the level of interest in a topic; thus, this correlation represents the extent to which editors in three different languages are responsive to relevant events that may require changes in the content of the pages. Figure 3 shows the correlation between the page views and edits on each article. In an ideal world, for those articles that are directly related to an event, the page would be updated on the same day as the event, since that is when people tend to flock to Wikipedia in droves, as we saw earlier in this section.

A1-EN ($r = 0.855$), and to a lesser extent A1-HI ($r = 0.538$), showed a positive correlation between the volume of edits and page views, as did A2-EN ($r = 0.738$) and A2-UR ($r = 0.616$). Consistently, English pages yielded the highest correlations for all topics, implying that more efforts were made to keep the English pages up to date than their Hindi and Urdu counterparts. This result also suggests that A1 (*Article 370*) and A2 (*Pulwama Attack*) are more directly connected to events than the other articles. By contrast, A3 (*Insurgency in J&K*) and A4 (*Kashmir conflict*), where the correlation is never greater than 0.4, can be perceived to cover rather constant topics; that is, ongoing, relevant, but not dynamic. However, A5 (*J&K Reorganization Act, 2019*), like A1 and A2, would seem to be as directly related to a recent event as an article can be — it explains the J&K Reorganization Act that was passed in August 2019 as a direct consequence of the revocation of Article 370. The correlation is still low for A5 because edits preceded views in all three languages. The event occurred, the article was created and edited without being viewed much or at all, and then people came to the page; in fact, for A5-UR, there were no more edits at all once people started visiting the page, whereas for A5-EN and A5-HI, there were a few more flurries of editing after that point. To summarize, recent events demanded more edits to A1 and A2 than to A3 or A4, and the English Wikipedia editors responded more swiftly than either Urdu or Hindi editors. We saw deceptively low correlations (or 'N/A,' in the case of A5-UR) on the A5 articles because the pages were *created* in response to the passing of the J&K Reorganization Act in 2019; once created, the information was fairly static, not requiring many edits.

4.3 Responsiveness to Malicious Edits or Mistakes

Another way to measure how active the Wikipedia community is in each language is by looking at reversion times; that is, how quickly editors make repair edits. These repair edits are often associated with undoing edits that harm the quality of the article in question, either because they are potentially malicious (e.g., vandalism, propaganda) or simply contain mistakes [5, 28]. We used 25^{th}-percentile (Q1), median, and 75^{th}-percentile (Q3) values as the measures of central tendency, since several articles had a few extremely long, outlying reversion times. The result showed that English editors reverted the most, and were the quickest to revert, whereas Hindi editors made fewer reversions and were sometimes slow to respond. Urdu editors reverted only rarely. See Table 3.

English editors acted quickly across the board, with 75% of reversions taking less than ten hours for all the articles in our corpus except for *Insurgency in J&K* (A3), and even there, the median time-to-revert was only three hours, with 75% of reversions taking about 40 hours or fewer. The Hindi articles had between 13 and 27 reversions each, and for response times, they were split. On the high end, for *Insurgency in J&K* and *Pulwama Attack*, the median reversion time was over two months. On the low end, *Article 370*, *Kashmir conflict*, and *Reorganization Act* had median response times of 9 hours 30 minutes, 1 hour 36 minutes, and 1 hour 24 minutes, respectively, with 75% of reversions taking less than a day for all three.

For Urdu, there were too few reversions to draw any conclusions about response times. Out of 255 edits made to the four Urdu articles in our corpus, only four were reverted. This could be attributed to a small good-faith editor count; however, based on the account of our Urdu interviewee and our observations about Urdu article content, we believe the extremely low number of reversions can be

Article	English				Hindi				Urdu			
	count	Reversion time (hrs)			count	Reversion time (hrs)			count	Reversion time (hrs)		
		25%	50%	75%		25%	50%	75%		25%	50%	75%
Article 370 (A1)	163	0.11	0.62	2.46	21	4.75	9.50	22.97	0	N/A	N/A	N/A
Pulwama Attack (A2)	42	0.09	0.45	0.50	13	19.54	1849.18	1849.18	2	No data[12]		
Insurgency in J&K (A3)	43	0.56	3.08	40.39	24	0.39	1717.76	3396.11	No page for Urdu			
Kashmir conflict (A4)	152	0.50	3.58	7.66	27	0.57	1.64	17.38	1	2.98	2.98	2.98
Reorganization (A5)	81	0.11	0.57	1.81	4	0.01	1.44	9.46	1	0.01	0.01	0.01

Table 3. Number of reversions and time-to-revert quartiles, in hours

attributed to the relative inactivity of the Urdu Wikipedia community and differing collaborative practices (see § 6.3).

Based on what we gathered from interviewees about malicious edits tending to be anonymous (IP edits) — whether because offending editors wished to be anonymous for nefarious or legitimate reasons [45], or because they were simply casual editors who had not made accounts — we decided to examine the ratio of IP edits to total edits in each article. IP edits are made by users who are not logged in, so their contribution is associated with their IP address rather than a username. The proportion of IP edits was largest on the Hindi articles (mean = 27.5%, std. dev. = 16.6); followed by English (10.5%, 9.5); followed by Urdu (4.9%, 5.8) — see Figure 2-(h) for the case of A1.

4.4 Editors Work Across Articles, but Rarely Across Languages

We checked if editors work across languages on the same topics, assuming that the more editors work across languages, the more the language editions can cross-pollinate. In other words, we hypothesized that there are benefits to both the smaller language editions and the English edition if editors work across languages. In one direction, big Wikipedia to small, we anticipated the benefits of large-scale participation (e.g., more eyes reading the news and potentially updating articles); in the other, the English Wikipedia might benefit from local news sources and perspectives that may not be readily accessible to the average English editor. Between the set of unique English editors of the articles in our corpus and the Hindi set, the Jaccard coefficient (intersection over union) was 0.014, and the Jaccard coefficient for English and Urdu was 0.003. Even though English is a major secondary language in both India and Pakistan, cross-language edits were rare.

However, these coefficients do not tell the whole story: the set of English editors dwarfs both the Hindi and Urdu sets by a factor of 5 in the first case and 24 in the second. Out of the 137 Hindi editors, eleven (8.0%) also edited one or more of the English articles in our corpus. Out of the 28 Urdu editors, two (7.1%) also edited one or more of the English articles in our corpus. Between Urdu and Hindi, the Jaccard coefficient was 0.012. It is worth noting that the lack of cross-language activity does not necessarily mean that Indian and Pakistani editors do not tend to edit the English Wikipedia. On the contrary, based on the locations of the IP addresses, we noticed that some of the edits to articles in English were made from India and Pakistan, and an interviewee confirmed that Indian and Pakistani editors frequently edit the English edition; in fact, many choose only to edit the English Wikipedia, it being the most widely read Wikipedia.

We found that many editors worked across articles in the same language. Figure 4 shows the relationship between the percentage of editors who worked on more than one article in the same language (y-axis) and the number of unique editors of the article (x-axis). With one exception, the Hindi *Reorganization Act* article (A5-HI), all the Urdu articles saw a higher proportion of editors who also edited one or more of the other articles in the same language than did the Hindi and English articles. The Hindi articles, in turn, tended to have a slightly higher proportion of

[12]This article had two reversions, but we were not able to get the reversion times with the `mwreverts` library.

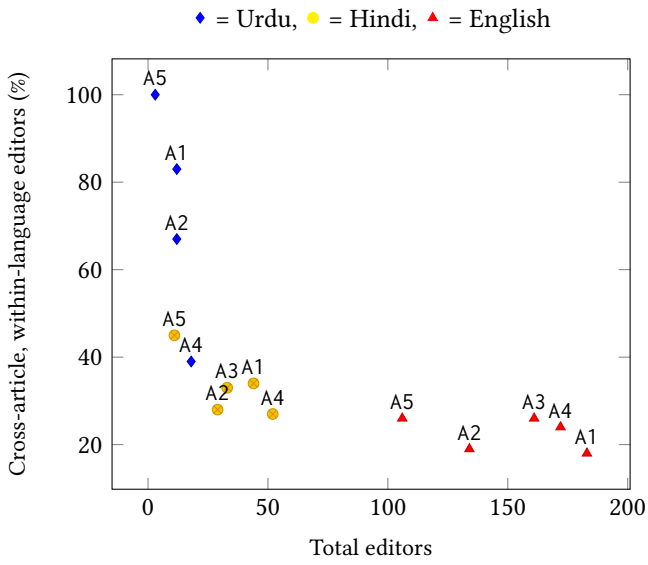

Fig. 4. How often editors worked across articles in the same language, versus editor pool size, per article

cross-article editors than the English articles. Within J&K-related articles, the size of the editing community seems inversely related to the amount of cross-article editing.

5 QUALITATIVE ANALYSIS ON (LACK OF) NPOV IN CONTENT

As there is no regulation of article content and structure across languages, it is reasonable to examine differences between English, Hindi, and Urdu articles. Moreover, in this situation, a majority of Urdu speakers are from Pakistan and a majority of Hindi speakers are from India. This could lead to differences in content, point of view (POV), and structure due to differing perspectives, as well as the scale difference of editor groups for the Hindi and Urdu Wikipedias (in comparison to the English Wikipedia). In the following sections, we discuss structural and content differences between articles in order to understand to what extent an NPOV might have been compromised in each community.

5.1 Structural Similarity

In general, the English pages were the most detailed, Hindi less so, and Urdu pages were often shorter than or equivalent to the Hindi pages, with the exception of the Hindi page being the shortest for A5. For example, in A1-EN, there are additional sections detailing presidential orders related to Article 370, the autonomy of J&K, the calls for Article 370's revocation, human rights violations, women's rights, and actions of the Union government of India, none of which are part of its Hindi and Urdu counterparts.

We found structural similarities for some pages: (A2-EN, A2-HI, A2-UR) and (A5-EN and A5-UR). Although structurally similar, the content on English pages was more detailed than on non-English pages. Most notably, A3-HI was a literal translation of most parts of the English article, with the section structure being almost a direct copy of the English article's structure. Overall, we found that the English articles are more structured and organized with sections and subsections, and their content more detailed, compared to the non-English articles. Some of the non-English articles seem to be modeled after English articles on the same topic.

However, even with similarities in structure, non-English versions are not mere subsets of the English version. The English article (A4-EN), for instance, devotes a section each to the Indian view and the Pakistani view of the dispute, demonstrating implementation of the NPOV guidance to give *due weight* to the parties in a dispute. In the meantime, the Urdu article (A4-UR) gives more space to *Kashmiri freedom fighters*[13] than do A4-HI or A4-EN. According to the NPOV policy, Wikipedia articles should "represent all significant viewpoints [...] in proportion to the prominence [14]." This exemplifies the challenges in evaluating *vNPOV* edits. For example, considering that many people in Kashmir speak and write Urdu, Urdu editors may have more detailed descriptions about the Kashmiri POV, or they may simply feel this POV is more prominent. While we cannot judge whether that weight is due or undue, and whether it constitutes *vNPOV* or not, the disparity of prominence between articles in different languages indicates tension between which views are given due or undue weight in the various language editions.

5.2 Partial Tones and Expression in Non-English Articles

Following NPOV policy guidelines (§ 1.2), we also analyzed whether articles reflect an impartial tone. We again highlight that simply presenting POVs is encouraged, as offering multiple perspectives is necessary in Wikipedia articles. However, the *Impartial tone* guideline requires editors to neither endorse nor reject a side. Additionally, according to the *Words to watch* guideline, editors should avoid "expressions of doubt" and other charged language [14]. Therefore, we analyzed the target articles to see if we could find expressions of a subjective character, presented without verifiable sources. We were able to easily find content that does not use an impartial tone in the non-English articles.

Frequently, we found that articles in one language contained information that, instead of presenting facts, presented the subjective opinion of one or more editors. The following quotes from two Hindi articles in our corpus are examples of such additions.

- A1-HI: *"The merger of Jammu and Kashmir into India was a much greater* ***necessity*** *and, to carry out this work, some special rights were given to the people of Kashmir at that time under Article 370."*
- A4-HI: *"The war exposed Pakistan's inadequate standards of military training, its officers of misguided selection, poor command and control systems, poor intelligence and bad intelligence procedures. Despite these shortcomings, the Pakistani Army managed to fight the larger Indian Army."*
- A4-HI: *"After this, Pakistan tried to occupy Kashmir in 1965 with its military force, due to which it had to face another* ***mouth-shattering defeat****."*
- A4-HI: *"The question arises that Jammu and Ladakh are happy with India since independence, but* ***why is Kashmir not happy****? However, experts say that at the moment, only 2% Kashmiris have come in Pakistan's* ***tricks****, all the rest love India."*
- A4-HI: *"The people of Kashmir on both sides, whether they are Muslims or non-Muslims, are leading lives of misery and pain. It is separatists, terrorists and their masters who are* ***having fun at their cost****. Their children roam around the country and abroad and live in all kinds of luxury. (...) The fundamentalist Kashmiris do not understand this and they do not even want to understand this."*
- A4-HI: *"Pakistan lost in this war. Stung by the defeat, Pakistan acted to spread hatred towards India and the entire politics of Pakistan became based on Kashmir, that is,* ***if you want power, then talk about occupying Kashmir****."*

[13]From the Indian POV, the Kashmiri freedom fighters are considered separatists, terrorists, and a proxy for Pakistan. The Pakistani POV sees them as freedom fighters, based on A4-EN.

It is not difficult to see that the above excerpts from the articles are not necessarily factual and favor one particular party. Indeed, most of the content in these articles does not have a corresponding reference and reads more as opinion — subjective and possibly even aggressive in its tone. Words like "stung," "tricks," and "mouth-shattering" in this context are plainly expressions that should be avoided according to the guidance.

Some of them are milder; for example, the *necessity* argument in A1-HI implies that merging the J&K region was beneficial for the people in the region, justifying the event from India's standpoint. Many of the examples come from *Kashmir conflict* (A4-HI), which is the most general article and covers other topics.

These quotes serve as evidence that an NPOV is poorly maintained in some Hindi articles. Notably, of the A1's, the Hindi article (A1-HI) had the highest IP edit ratio (§ 4.3), with 27.5%, almost three times the IP edits as the next highest (A1-EN, with 10.5%). This might indicate that IP editors tend to contribute *vNPOV* content. While our analysis cannot speak to this directly, it aligns with several interviewees' perceptions of IP edits. We will return to this in more detail in § 6.4.

More mildly opinionated views emerged in both Urdu and Hindi articles. While Hindi editors see the conflict in Kashmir as a dispute between India and Pakistan only, Urdu editors' views on the conflict seemed to involve another group, consistently stating there is a third group in the Kashmir equation.

- A4-HI: *"Kashmir dispute is going on between India and Pakistan since 1947 over the authority on Kashmir. For this, Pakistan had attacked thrice on India and all three times it was severely defeated."*
- A4-UR: *"The Kashmir issue is a dispute between Pakistan, India, and* ***Kashmiri freedom fighters*** *over the ownership of occupied Kashmir."*
- A4-UR: *"Some of the groups of freedom fighters of Kashmir are in favour of complete autonomy of Kashmir while some others want it to be a part of Pakistan. India claims complete ownership over Kashmir."*
- A4-EN: *"The Kashmir conflict is a territorial conflict* ***primarily*** *between India and Pakistan over the Kashmir region."*
- A4-EN: [from the section 'Indian View'] *"Insurgency and terrorism in Kashmir is deliberately fuelled by Pakistan to create instability in the region. The Government of India has repeatedly accused Pakistan of waging a proxy war [Kashmiri separatist being a proxy of Pakistan]."*
- A4-EN: [from the section 'Pakistani View'] *"Pakistan suggests that this means that Kashmir either wants to be with Pakistan or independent."*

Urdu editors' perspectives differ subtly from that of Hindi editors.

We presented a few examples of *vNPOV* edits that appeared in our target articles. While there are many reasons this *vNPOV* content may have been created — for example, its contributors may have been politically motivated, or they may have been unfamiliar with or not understood the NPOV guidelines — the fact remains that the *vNPOV* comments that we presented remained in the articles at least until the time of this study's submission. Interestingly, our Hindi translator left a personal note that they were "surprised" by how different the Hindi articles were from their English counterparts, saying the Hindi articles "can be termed as a bit biased." It is certainly good evidence of how challenging it can be to maintain an NPOV on a controversial subject.

5.3 Discrepancies in Information

It was not difficult to find discrepancies in information among articles. Discrepancies ranged from contradicting facts to certain information being present in only some of the articles. For the *Pulwama Attack* articles (A2), there are disagreements between articles on the facts. The most

notable example of this is in A2-HI, which gives two different numbers of CRPF members killed in the attack.

- A2-EN: *"The attack resulted in the deaths of **40** Central Reserve Police Force (CRPF) personnel and the attacker."*
- A2-HI: *"On February 14th, 2019, a convoy of CRPF vehicles carrying Indian Security personnel was attacked by Pakistani terrorists on Jammu Srinagar National Highway, in which **45** security personnel [CRFP] lost their lives."*
- A2-HI: *"On Thursday, a convoy of Central Reserve Police Force [CRPF] Indian security personnel was attacked by militants in Avantipora area of Jammu and Kashmir's Pulwama district, in which about **40** soldiers have lost their lives and many others have been injured."*
- A2-UR: *"As a result of this attack the terrorist and **46** members of Central Reserve Police Force CRPF were killed. An separatist organization named Jaish-e-Mohammed accepted the responsibility of this attack."*

Although it seems that the discrepancy is coming from editors using different sources that may have had differing information immediately following the event[14], it seems that little effort had been made to update the articles as more accurate information became available.

In another case, the discrepancy seems more deliberate. In the same article (A2), each version attributes the attack to a different group. Here are direct quotes from each article.

- A2-EN: *"The responsibility for the attack was claimed by the Pakistan-based Islamist militant group Jaish-e-Mohammed. The attacker was Adil Ahmad Dar, a local from Pulwama district, and a member of Jaish-e-Mohammed. India has blamed Pakistan for the attack. Pakistan condemned the attack and denied any connection to it."*
- A2-HI: *"On February 14th, 2019, a convoy of CRPF vehicles carrying Indian Security personnel was attacked by **Pakistani terrorists** on Jammu Srinagar National Highway, in which 45 security personnel lost their lives."*
- A2-UR: *"The Pulwama attack was a **terrorist attack** on a convoy of Indian security forces on February 14, 2019."*

The English article not only explicitly revealed the location and name of the terrorist group involved in the attack, but also introduced what each country claimed. The Hindi article called the attackers Pakistani terrorists, with no nuance, while the Urdu version called the attack a terrorist attack, omitting where the terrorists were from. It seems that the Hindi and Urdu versions partially hid information that would have given more context for the attack in favor of national interests. Here is another, similar example that shows a clear difference in how the Urdu, English, and Hindi articles describe Indo-Pakistani wars as part of the *Kashmir Conflict* article (A4).

- A4-EN: *"Pakistan claimed that the captured men were Kashmiri '**freedom fighters**', a claim contradicted by the international media."*
- A4-HI: *"Kashmir dispute is going on between India and Pakistan since 1947 over the authority on Kashmir. For this, **Pakistan** had attacked thrice on India and all three times it was severely defeated."*
- A4-UR: *"The Kashmir issue is a dispute between Pakistan, India and **Kashmiri freedom fighters** [کشمیری حریت] over the ownership of occupied Kashmir."*

A4-UR defines the conflict as occurring between Pakistan, India, and *Kashmiri freedom fighters* within its text. By contrast, the A4-EN article reflects Pakistan's claim that Kashmiri freedom fighters are a separate group from the Pakistani army, but states that this claim has been "contradicted by international media." A4-HI defined it as a conflict just between India and Pakistan, never using the

[14] According to many one-year anniversary articles, 40 CRFPs were killed [33].

term "Kashmiri freedom fighters," instead deeming them "Pakistani terrorists." In both cases, the English version provided balanced, more neutral information than the two non-English articles.

In some cases, important pieces of information that are closely related to a topic are missing or only vaguely described. A1-UR describes the provisions of Article 370 with minimal information, while A1-HI and A1-EN are more detailed. It is challenging to gauge the particular intent behind this discrepancy, but it seems that Article 370 is of less interest to Urdu editors than to Hindi and English editors. The page length of A1-UR, compared to the other two, supports this idea as well. Consistently, in A1-UR, there is no mention of the revocation of Article 370, which is a significant event to the topic (see the page view increase immediately after the event in Figure 2).

Based on our content analysis result, fortunately, we did not find the English articles to be problematic in general. While there can be a negative implication in information flowing from one language to another Wikipedia, we conclude that cross-checking content over multilingual articles can be helpful to reveal potential discrepancies and resolve conflicting information.

6 INTERVIEW STUDY: UNDERSTANDING EDITORS' GOALS AND PRACTICES

To better understand wiki editing practices, we interviewed six editors who participated in the target articles (A1–A5). Key themes were extracted by thematic analysis of the interview transcripts and interviewer notes. All the interviewees participated in one (or more) target articles. Their demographic information is specified in Table 1. Three of them (E1-US, U1-PK, E2-IN) are administrators, and one (E3-SB) is a project administrator, which indicates that they have had long experience with Wikipedia, not only in editing, but also in acting as stewards and arbiters. The following section discusses our understanding of their Wikipedia practices. We categorized each theme into motivations, communication, writing process, and NPOV.

6.1 Motivation: Hobby of Being a Good Global (and Local) Citizen

We wanted to understand editors' motivations for participating in Wikipedia articles in general, to see if any difference among editors emerges or to see if anyone has a particular interest in this topic. In response to the question asking about their motivation for editing Wikipedia articles, all interviewees expressed an interest in improving the quality of content and size of Wikipedia content in general. Rather than being specifically motivated by one particular topic, this motivation was sparked by a personal interest and has since become a "passion" or "hobby." They also shared a sense of contribution in participating in Wikipedia and making correct information accessible to the broader population. The following quotes represent their altruistic motivations well.

- E1-US: *"It's my hobby. So it's a thing that I find enjoyable, it's relaxing, it's a thing I can do to kind of unwind or a thing I can do in my free time. (...) But I like the fact that this hobby allows me to help others to share knowledge to, you know, express my belief in the value of free culture. (...) I think that the world's knowledge should be free to everyone and that includes persons with certain kinds of different abilities or disabilities."*
- U1-PK: *"I started using it because I had some knowledge/information that was not present or was incorrect on Wikipedia. So to correct that, I wanted to make a contribution."*

Two of the interviewees who were non-English editors also mentioned contributing to the Wikipedia community in their own language, especially when they see a lack of information in their language.

- H2-IN: *"I saw that there was a dearth of material in Hindi, that is why I thought I should contribute."*

- U1-PK: "*So mostly when these Wikipedia pages related to my profession come to me*[15]*, I translate and update them to Wiki.*"

Our Hindi and Urdu editor interviewees described a passion for spreading the articles in their languages. They hoped to bring a more polished and mature voice, similar to that of the English edition of Wikipedia, to the smaller or nascent language editions.

All interviewees except H2-IN were interested in contributing to Wikipedia articles in multiple languages. Their editing interests range from "simple maintenance work," which does not require them to be fluent in the language, to learning new languages to be able to contribute to the articles.

- E1-US: "*I can kind of tell like, this doesn't seem quite right, even though I don't really know what they're saying. So there were some things missing. And I tried to encourage others to include those things. So that's one thing I'll do for other languages that I don't really have confidence in. (...) I'll just add images, you know, I don't necessarily have to know a lot in those languages to be able to add some kind of media or add something.*"
- E2-IN: "*I do some technical work with other Wikipedias, such as Bengali (and recently Corsican)*"
- U1-PK: "*I'm learning Hindi, I've learned a lot. I wanna contribute to Hindi as well but the obstacle is the script Devanagari. So I want to contribute, but I am learning or looking for typing software to type in Hindi. So I would love to contribute in Hindi too.*"

Some would do simple tasks, such as fixing typos ("housekeeping") or formatting edits, while H1-IN is incubating a Wikipedia community for an entire language, Marwari.

None of the interviewees were particularly interested in the Jammu & Kashmir topic. Instead, they were interested in a wide range of topics, from music to cricket players [16]. In fact, many of the interviewees did not even remember editing any of the target articles. Participating in various types of topics is a good indicator that an editor has good intentions, as opposed to those who have a specific agenda to convey in Wikipedia — so-called "single purpose accounts" (SPAs), which "end up getting blocked for one reason or another"(E2-IN) [12]. Rather, they perceived Wikipedia editing as a "passion" or "hobby" that they engage in without political motivations on any topic. The following quote from E3-SB represents their attitude towards Wikipedia articles well.

- E3-SB: *"I have my own personal views on many issues, like everyone. I cannot say I don't have them. However, if things are to be recorded properly, then one has to put aside his personal feelings are present only the facts."*

In addition, all interviewees had a sense of contribution in generating knowledge accessible to a broader audience, in line with the goal of Wikipedia.

6.2 Process: How They "Present Only the Facts"

In their writing practice, it seemed that our interviewees were committed to generating accurate and neutral content. On the other hand, editors that are proximate to regional conflict topics may exhibit self-focus bias [22, 43]. When we asked our interviewees about what sources they used when editing, we found that all Hindi and Urdu editors used English Wikipedia articles (or the references therein) as their primary sources, sometimes directly translating material from an English page to its counterpart on the Hindi or Urdu Wikipedia. They also mentioned that translation is a primary activity within their participation in Wikipedia.

- H1-IN: "*English Wikipedia, I use the English version and their sources as well.*"

[15]Interviewee U1-PK is an online news editor, and he reads English wiki articles and translates them as part of his job.
[16]E3-SB added that he is particularly interested in articles about military conflicts, sharing his personal story that one of his family members was killed in action. He added that he recently participated in the article on Turkish military operations in Syria.

- U1-PK: "*I think to myself, 'if this is not present on Urdu Wiki, and is present on English Wiki, then I take a couple of paragraphs from my page/my website article and create a new page/edit for the Urdu Wikipedia and then I cite accordingly.*"
- H2-IN: "*When it comes to Hindi, where most of my edits are its mostly content creation through translation.*"

The English editors, usually sourced information from known "reliable" sources, such as reputable news organizations, public databases, journals, or scholarly articles.

- E1-US: "*I did a literature review, using JSTOR, and finding scholarly sources. And then there are some, I think I included some kind of popular level sources there but that included some proper academics and that's the sort of thing that requires that.*"
- E2-IN: "*We maintain a list of well-known sources as well, and have discussions on whether they are reliable or not, depending on the consensus, we can choose to either mark it as reliable, or unreliable or blacklist the source as well.*"
- E3-SB: "*I personally tend to use the more neutral sources like AFP, AP, Reuters, etc. (...) If neutral sources don't exist I use non-neutral sources, but properly attributed the info to the source and let the reader decide on their own if they trust the information or not. (...) Wikipedia already has a list of sources that are considered reliable, like the ones I mentioned.*"

It seems that all the interviewees were cautious about the sources they choose. Although biased sources are not strictly forbidden on Wikipedia — it is how opinions are presented and attributed that is important to NPOV [14] — it seems that all the interviewees have taken care to systematically avoid sources that may propagate biased perspectives, when possible.

6.3 Collaboration: Off-site Communication for Non-English Editors

To understand how editors in different languages collaborate and coordinate differently, we asked how they resolve conflicts and communicate with other editors. While the most visible collaborative interactions happened in the talk or user pages, some interviewees said that they collaborate outside Wikipedia (off-site communication). Some South Asian editors (H1-IN, U1-PK, E2-IN) mentioned that they used various virtual and offline channels to communicate with one another informally, delegate work, and discuss issues.

- U1-PK: "*Yes, we communicate frequently. On the platform, Wiki as well as other means as there are WhatsApp groups by certain admins/more active users. And we interact with them all. (...) Sometimes I doubt whether the information is true on the page, so I ask the editors to add sources. What we do on Facebook, is ask if they wanna edit a certain page or not. Or we delegate it to someone.*"
- H1-IN: "*We use [talk pages] when we want to have a talk with an editor but you don't have a way of communicating with them such as Facebook or WhatsApp.*"
- E2-IN: "*Probably via off-wiki means, WhatsApp or the likes, [WhatsApp]'s common across the world but particularly for South Asian regions.*"

These interviewees shared various types of off-site means of communication (e.g., Facebook groups, Facebook Messenger, SMS messages, WhatsApp, IRC, Zulipchat). U1-PK also noted, responding to a question about how editors know if another editor is reliable or not, that the community is small enough for him to know by the screen names:

- U1-PK: "*Some are the ones with names or the ones you already recognize. Or a new name or new users or new IP [addresses]. And there are a lot of users in English, but with the Urdu editors, you can tell when there's a new name. (...) Any controversial edit must have a reference, or else it gets deleted. Or normal talks get unnoticed if it is a reputed editor.*"

U1-PK told us that he is an administrator [9] on the Urdu Wikipedia, and he says that the Urdu community is small enough for him that he can tell whether an editor is new or experienced ("*There might be like 100 [wiki editors] or so from Pakistan.*"). It seems that Wikipedia editors form a small, but strong community that is connected outside Wikipedia. This can be efficient on a small scale, but there may be negative implications in collaborating off-site, as discussion and coordination are not archived as part of the article history. This may result in the loss of context, preventing further research and data analysis.

6.4 NPOV: Dedicated to Neutrality

General differences in opinion on content or references cause conflict between editors, but conflict also arises because of *vNPOV* edits and intentional vandalism. All the interviewees confirmed that they strive to adhere to Wikipedia's NPOV policy [14]: presenting information impartially, citing reliable sources, and watching out for problematic edits. Our interviewees confirmed how dedicated they are to sustaining neutrality.

- U1-PK: "*No, we always stay neutral ... we know that if we write as per our bias, some Indians will read it and change/correct it and this can always escalate. So we try to be neutral and always with a citation so no one can say a thing*".
- E3-SB: "*So the articles would be historically as precise as possible, since Wikipedia is considered a form of online encyclopedia. And by 'precise' I mean to present things as they are, from a third party point of view, without any personal non-neutral editing getting involved.*"

We also asked the questions about whether they have witnessed any kinds of *vNPOV* edits, including vandalism. All interviewees described such *vNPOV* edits occurring to varying degrees. Here, we introduce a few of their anecdotes that they shared from the past:

- U1-PK: "*There may be political vandalism. Political vandalism where two parties keep vandalizing their pages, but there are Pakistani certain vigilantes who protect their pages.*"
- E1-US: "*There is one [vandal] who's a cross-Wiki vandal (...) What he will do is to post a picture of a chimpanzee and he'll put a little caption underneath that says it's Omar al Bashir [former president of Sudan].*"

We also asked if they felt any tension between India and Pakistan while editing.

- U1-PK: "*No no, there aren't a lot of biased edits. We remove it if it is unreferenced. And we are personally connected with a lot of Indians or groups. So we try to not escalate the matter and pay attention to citations.*"
- H1-IN: "*Yeah, I'd say that the moderation isn't that good. Some pages are quite biased. (...) there are people whose articles are so politically biased they aren't objective anymore.*"
- H1-IN: "*Yes, in Hindi Wikipedia, I see a lot. I suspect English Wikipedia does not have a lot.*"
- E2-IN: "*I think there has been [vNPOV edits], yes. Or at least there was. (...) There would be something like 'J&K belongs to India' or 'congress did this' kind of unwanted contribution. But it was less muted than other topics*"
- E3-SB: "*Very often. Especially if it's a current ongoing subject. People like to either edit in their own personal POVs. Or make edits based on biased sources. (...) What I can personally tell you, when looking at the Serbian, Croatian or Bosnian Wikipedia's, differences exist. And they again mostly come from the different POVs of the people.*"

One anecdote we heard from an English editor in India (E2-IN) helped us realize that sustaining an NPOV is challenging, not simply because of the need to moderate *vNPOV* edits, but also because of actual threats, depending on the political environment that editors live in.

- E2-IN: *"Recently a colleague of mine was doxxed by right-wing media outlets and IPS (Indian Police Service) officer*[17]*, leading him to be outed and vanishing his account, it was quite a dangerous time to be an editing that particular article [in English], resulting in most Indian editors just stepping away, including me, from editing contentious topics. Eventually, the article was locked, the talk page was locked - that's very rare - to stop the influx of clearly POV editors who wanted to bias the article. (...) The editor I was talking about was threatened with arrested by a public servant. (...) Another colleague of mine had threats sent to his family."*

The fact that people in power in India pay attention to English Wikipedia articles on this topic, and feel the need to intimidate editors, is concerning, but also indicates that Wikipedia content is significant enough to warrant attention from certain political parties

The interviewees also shared their experiences with how vandalism or *vNPOV* edits typically emerge. Largely, we found two different pattern: joint action and anonymous edits.

- E2-IN: "*It is general knowledge that on contentious articles, there are groups of such accounts that become active and randomly start editing in concert.*"
- E3-SB: "*Between, I would like to add that there have been several instances during the years where if a large enough number of biased editors show up on an issue then Wikipedia guidelines are ignored and their own view of the issue is cemented in an article.*"
- H1-IN: "*Such people are the ones who use IP addresses. People who use userID's, they are trustworthy. If a person does a lot of vandalism, that person is blacklisted.*"
- E3-SB: "*The more unreasonable editors tend to be unregistered IP editors who are there in my opinion only to push their own side's POV.*"
- U1-PK: "*We specifically look at IP based edits [when we scan for potentially problematic edits]*"
- E2-IN: "*Contrary to popular belief, most IP addresses make good edits, as do most accounts. I'm talking by percentage of total edits here.*"

However, we were told that conflicts do not last long. Conflicts are largely resolved through wars of attrition (editors who stand their ground on the pages for longer tend to get their way) or Wikipedia administrators banning such users.

- E3-SB: "*Through either their talk pages or the discussion pages of the articles that are in dispute. And I tend to point out Wikipedia's policy and guidelines and try and find a compromise solution. (...) If most people agree that a proposed solution is reasonable and according to the guidelines then they overrule those that do not want to compromise. Those kinds of people tend to continue edit warring which can get them blocked on Wikipedia.*"
- E1-US: "*When there's a conflict, a lot of times the way that gets resolved is really just through attrition, right? (...) And the one who really gets his way is the one who kind of waits the longest or who...you know, sometimes you can game the system, but sometimes it's just like the other person gets exhausted, and in reality that's how it works.*"
- H1-IN: "*I have also seen old information about an admin that used to have more than one account and he used the other account for vandalism. I don't know how he got caught, but then he was blocked.*"

In general, most of our interviewees are moderately confident that NPOV is well-maintained in Wikipedia articles, even with many potential risks involved.

- E2-IN:"*Firstly vandalism is dealt with using* `ClueBot NG`*, it's a quite popular bot that has a huge database of what are vandalism edits like, so very obvious and sometimes less obvious vandalism is caught and reverted by the bot, the bot also warns them appropriately.*"

[17] A right-wing party (the Bharatiya Janata Party) was the ruling political party in India at the time of submission.

There exist many automated anti-vandalism tools and active research on vandalism detection [1, 38, 40]. However, much of this research is focused on the English Wikipedia, and some of the tools cannot be readily applied to other Wikipedias. The Hindi Wikipedia community is large enough, in terms of consumption, to become vulnerable to *vNPOV* edits, but lacks the editor power to develop resources — automated tools and organized methods — for handling such edits, whereas the Urdu Wikipedia community, as of this writing, appears to have a small enough readership that adversarial editors are not motivated to vandalize.

7 DISCUSSION

As of this writing, the Hindi and Urdu Wikipedia communities are trying to emulate their English counterpart, but they are orders of magnitude away from matching the English Wikipedia in content, editor participation, and readership. The Hindi and Urdu communities are motivated by growing their language edition by adding more content. Their editing practices reflect this — they seem to focus more on adding content than on the appearance, organization, or structure of the articles. That being said, there is a gap between the intent that drives their efforts and the impact it achieves, due to their smaller size (see § 4.1). Our Urdu interviewee (U1-PK) especially seemed to know everyone in his editing cohort. When new people join the editing force, they are easily noticed, and may be "trailed" or mentored to suss out their motives.

7.1 Given Enough Editors, All Issues Are Noticed

The more eyeballs there are on a page, the more editors — and by extension, the more people with differing points of view — will be moved to edit. This is akin to *Linus's Law*[18] in software development: "given enough eyeballs, all bugs are shallow." In this context, "bugs" are missing or inaccurate information in an article. With more diverse points of view, articles naturally become more detailed and comprehensive. Inversely, if editors mostly share the same opinions, "bugs" in the article may go unnoticed. In the Hindi and Urdu editing communities, many editors know each other and talk off-site. When there is a need to discuss a detail in an article, they use talk pages much less than English editors do; consequently, there is no paper trail and little accountability. The English articles in our corpus, having more unique editors than any of the Hindi or Urdu articles, naturally presented a more balanced story, but with great power comes great responsibility [50] (or, with many eyeballs come many potential vandals).

7.1.1 Tension between well-intentioned growth and NPOV policies. In the course of our interviews, we were surprised that none of our editors were specifically invested in the articles in our corpus — some had no memory of making edits to any articles related to J&K. This may have partly to do with our recruitment criteria (we invited editors who had made *more edits* first). As Bryant et al. found, as editors gain experience, they turn their focus from specific articles to the good of the Wikipedia as a whole. We certainly found that to be the case with our interviewees, half of whom were not only editors but admin on their respective Wikipedias (U1-PK, E1-US, E2-IN). Others [35] have shown that those same, relatively few, editors who focus on the health of Wikipedia more broadly produce most of the content on Wikipedia. We see this further reflected in the Hindi and Urdu editions, where small numbers of editors who are devoted to growing their respective editions and presenting information as neutrally as possible. However, as we touched on above, these communities may be too small to be perfectly neutral. They may miss or misrepresent information simply by virtue of their small numbers, and an article may appear *vNPOV* when compared to its English cousin because of those omissions. In short, our results suggest there may be a tension between a language edition's ability to grow, and its ability to adhere to NPOV goals. We believe

[18] https://en.wikipedia.org/wiki/Linus%27s_law

the discrepancies we saw between the same article in different languages (§ 5) were mainly due to this kind of inevitable oversight. Truly malicious edits were usually dispatched quickly.

7.1.2 Exploiting data voids. There is a danger with polarizing topics, like J&K, of adversarial actors planting false or misleading information, especially when news has just broken and reliable sources may not have had time to react. In these times, adversarial actors exploit the "data void:" the absence of high-quality information about the new event [20]. Perhaps the one case we observed where a whole article was blatantly *vNPOV* (A3-HI) is an example of a data void being exploited: *Insurgency in Jammu and Kashmir* does not appear to be a frequently sought topic, usually garnering under a hundred views per day in the Hindi edition (fewer on average in 2019 than any other Hindi article in our corpus), so self-interested editors saw an opportunity to manipulate web search results for Hindi speakers.

On the whole, Wikipedia editors tend to be very responsive to real-world events. These responsive edits tend to be made by good-faith editors, and the ones that are not are quickly reverted (see § 3). Thus, we see evidence that the possibility of "natural" data voids is relatively low. That said, there is a temporal problem: readers are visiting these pages while they are still in flux and may contain *vNPOV* edits. It might be advantageous for Wikipedia to institute a new kind of tag or warning to appear at the top of a brand-new article, or section of an article, that has suddenly experienced a flurry of edits — akin to the NPOV dispute tags[19] — something to the effect of: *This is a hot topic that is still undergoing edits; read with caution, knowing that information may be missing or misleading.*

7.2 Challenges of Defining and Upholding Neutrality

The Wikipedia policy of NPOV dictates that articles explain the perspectives on a conflict like J&K "fairly and without editorial bias." More than that, the policy demands that perspectives be given space in an article *proportional to their prevalence in "published, reliable sources"* [14].

7.2.1 Reconciling Interviewee's "Bias" with NPOV Violations. In the course of our interviews, we noticed that interviewees often used the term "NPOV" when responding to our question about bias. When they would use the word "biased," it was generally to describe edits that used charged language or stated opinions as facts, or sources they did not deem reliable. The English interviewees told us that editors maintain a list of well-known sources by consensus, and will blacklist news sources, as well as individual editors, that consistently violate NPOV. According to our Hindi and Urdu interviewees, the smaller Wikipedias tend to use the same sources on an article that the English Wikipedia uses, except in certain circumstances, like when English-language news outlets do not cover an event at all, or not in as much detail as a local-language source.

In all three Wikipedias, based on our interviews and quantitative analyses, it is common practice with these controversial pages to revert *vNPOV* edits outright if they do not have a citation (U1-PK). In articles like A4-EN that are already "too long to read and navigate comfortably" (*Article size* tag[20] on the page at the time of this writing), perhaps this is a reasonable policy. The NPOV policy states: "Remove [sourced] material only where you have a good reason to believe it misinforms or misleads readers *in ways that cannot be addressed by rewriting the passage* [emphasis added]" [14]. However, perhaps, even when an addition does not provide a citation, the better course of action for an editor with a mind to uphold NPOV would be to seek out a source and rewrite the passage; that is, if the addition represents an important perspective that warrants its *due weight* in the article.

[19] https://en.wikipedia.org/wiki/Wikipedia:NPOV_dispute#Adding_a_tag_to_a_page
[20] https://en.wikipedia.org/wiki/Wikipedia:Article_size

This would serve the dual purpose of balancing the article and plugging a data void that malicious (or simply less-seasoned) editors would otherwise keep trying to fill.

7.2.2 vNPOV By Omission. The main manifestation of *vNPOV* that we saw in the articles was the absence of certain important details (violation of *Due and undue weight*); only in one case, A4-HI, did we see considerable blatant impartial tone. The Hindi and Urdu articles tended to be much shorter than their English counterparts and give different impressions of the conflict. For example, A4-UR characterized the conflict as between Pakistan, India, and Kashmiri freedom fighters, whereas the Hindi and English articles (A4-HI, A4-EN) did not describe any Kashmiri group as a third party to the conflict. According to A4-EN, it is the "Indian View" that Kashmiri separatists are "a proxy of Pakistan," and the "Pakistani View" that Kashmiri freedom fighters are divided on whether being independent or being part of Pakistan would be preferable (but either would be better than the current situation). The synopses offered by A4-EN of the Pakistani and Indian views pretty well matched the angles of the Urdu and Hindi articles (A4-UR, A4-HI), respectively; i.e. A4-UR presented the Pakistani POV and A4-HI presented the Indian POV.

7.3 The Arc of the Hindi and Urdu Editions Bends Toward Resilience

Our findings are a testament to the power of Wikipedia to influence public opinion. It has earned its reputation as a reliable, encyclopedic source — and the go-to place for information on all branches of knowledge, not least breaking news and such polarizing topics as Black Lives Matter [46] and the status of J&K.

The very principle of NPOV that makes Wikipedia a reliable asset also makes it a target for adversarial editors. With such high readership and a worldwide reputation as a source of trustworthy and neutral information, it takes little effort for an editor with malicious intent to add information to a given Wikipedia page and potentially influence public opinion. The well-meaning regular editors have to deal with vandals who make *vNPOV* edits (§ 1.2), e.g. by painting one or another party in the conflict unfairly, deleting information, or presenting opinions as if they were facts. In our study, the English articles (which garnered the most views by far of the articles in our corpus) drew the attention of the Indian Police Service (IPS) and perhaps other state actors (§ 6.4). However, attempts by nefarious actors to change the story, whether by intimidation or by edits, were no match for the combined force of seasoned Wikipedians.

Even with editors' commitment to NPOV, *vNPOV* edits still survive in some target articles (see § 5). We foresee that as the Hindi and Urdu editions grow, they will need to adopt more sophisticated or rigorous methods for managing and tracking conflict, as Kittur et al. observed in the early 2000s, when the the English Wikipedia was relatively new [28]. At present, Hindi and Urdu editors prefer to use WhatsApp and the like to coordinate with their cohorts (see § 6.3).

However, for a language with a larger corpus, like English, identifying and dealing with problematic edits is highly labor-intensive, so administrators require the assistance of automated tools. Tools for vandalism detection have been the subject of significant research in recent years [1, 38, 40]. Most researchers focus on the tools available in the English edition of Wikipedia. While some tools do exist for non-English editions, they are fewer and have received little attention from the research community. As these communities grow, we anticipate that there will also be a substantial increase in the number of edits. Thus, we see developing automated tools for anti-vandalism and that support NPOV policies within these smaller language editions as both (a) aligned with the goals articulated by our participants, and (b) a fruitful direction of future work.

8 LIMITATIONS AND FUTURE WORK

While the revocation of Article 370 was an important and relevant event, the Wikipedia articles analyzed herein may not be representative of other controversial events. Further studies could look into other events to better understand how articles and editor behavior differ across languages. Moreover, the articles were selected beginning with the English set of articles; the Hindi and Urdu articles were selected to match this set. There may have been more Hindi or Urdu articles relevant to the study that we missed because of our article selection process.

In the case of this study, six Wikipedia editors were interviewed across three languages. Interviewing them gave us valuable insights into their working habits, as well as their interactions on Wikipedia. Further studies could involve interviewing more authors across different topics, languages and locations. Our interviewees were by no means a random sample; self-selection bias was unavoidable. Additionally, we only interviewed editors with user accounts, excluding editors who edited anonymously (under an IP address only). This skewed our sample a bit more, albeit inevitably.

Our work here did not consider the persistence of edits, instead only looking at the number of reversions that occurred. While it is possible these reversions were undone at a later point, we do not consider this likely, given our analysis of the translated articles. However, we see a specific analysis of the persistence of edits focused around particular politically salient events as an important direction of future work. This would likely incorporate a metric similar to the PWR (persistent word revision) metric used by Panciera et al. [36], but it may need to be adapted slightly for the Hindi and Urdu Wikipedias, where the threshold number of revisions beyond which an edit is almost never reverted may be different[21].

Also, since a large portion of our work was based on analyzing interviews and articles written and spoken in Hindi and Urdu, it is possible that certain nuances in language structure and meaning were "lost in translation," but this drawback is inherent to all translations. That said, two of the authors of this paper are native Hindi speakers, and Urdu and Hindi are mutually intelligible, being varieties of a single Hindustani language.

Another opportunity for future work is the continued rollout of the Wikimedia Foundation's ORES project, which seeks to support and build tools for antivandalism work on Wikimedia infrastructure. ORES may be able to substantially address the issues of growth in the Hindi and Urdu language editions that we document here. However, we see an important direction of future work focused on exploring the feasibility of machine learning tools like ORES (which require large amounts of training data) to serve smaller editor communities, like those in the Hindi and Urdu language editions.

As we described, IP editors are those who are not logged in, such that only their IP addresses are associated with their contributions (see § 4.3 and 6.4). While Tran et al. [45] discuss IP editor behavior in the English language edition, examining the extent to which their research is replicable on non-English language editions of the project would enrich this area of research. Three interviewees told us that most vandalism is committed by IP editors, and two speculated that some of these editors might be paid propagandists. These paid propagandists are a growing threat, not only to Wikipedia or in the region we chose for this study, but to websites and organizations around the globe. On the other hand, one editor said that "[c]ontrary to popular belief, most IP addresses make good edits, as do most accounts," which aligns with the findings of Tran et al. [45].

Other directions to explore would be to study, broadly, how new language editions of Wikipedia are developed; how "house rules" are agreed on (or not); and why some projects fail (whether due to propagandists outnumbering regular editors or other factors), as the Bulgarian Wikinews

[21] Panciera et al., studying the English Wikipedia, set the threshold at five revisions.

project recently did [11]. With respect to house rules, we would like to know how the policies of NPOV, verifiability, and "no original research" vary or are interpreted differently in other language editions of Wikipedia.

9 CONCLUSION

Free access to unbiased information is extremely powerful, and consequently attracts the attention of POV-motivated groups or individuals seeking to influence public opinion. Despite efforts by adversarial actors, the dedicated editors from all three language editions in this study all strove to uphold a neutral point of view (NPOV).

All the editors we interviewed shared a passion for Wikipedia and considered themselves stewards of neutral information on this unique peer production platform — making accurate information available, keeping articles current, and, in the case of the Hindi and Urdu editors, growing their language edition (§ 6.1). Despite editors' best efforts, NPOV was hard to achieve for the Hindi and Urdu editions by virtue of the editing communities being so small (§ 7.1). The response times to new events and problematic edits alike were, on the whole, quick, although the Urdu data were sometimes inconclusive, and there were exceptions among the Hindi articles (§ 4.2 and 4.3). The growing army of volunteer contributors is a testament to the success of the Wikimedia Foundation's model. As the Hindi and Urdu Wikipedia communities grow, we anticipate that their respective editions will continue on their trajectory and become ever more reliable sources of unbiased information for readers around the globe, even for such polarizing topics as Jammu & Kashmir.

ACKNOWLEDGMENTS

To Aaron Halfaker, who developed the *mwreverts* library; to Cory Klingsporn, whose outstanding proofreading raised the caliber of this publication; and to all the editors who spend their free time making Wikipedia a reliable source of neutral information for people around the globe. Thanks also to Hally Sablosky for her tremendous support in the eleventh hour. We also thank the students in CS5734: Social Computing and Computer-supported Cooperative Work (Fall 2019) at Virginia Tech for providing feedback during in-class presentations. We are grateful to Engineering Faculty Organization (EFO) at Virginia Tech for supporting this work.

REFERENCES

[1] B. Thomas Adler, Luca de Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. 2011. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In *Computational Linguistics and Intelligent Text Processing*, Alexander Gelbukh (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 277–288.

[2] K Louise Barriball and Alison While. 1994. Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing-Institutional Subscription* 19, 2 (1994), 328–335.

[3] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[4] Susan L. Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/1099203.1099205

[5] Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. 2006. Temporal Analysis of the Wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI '06)*. IEEE Computer Society, USA, 45–51. https://doi.org/10.1109/WI.2006.164

[6] Xiaoxi Chelsy Xie, Isaac Johnson, and Anne Gomez. 2019. Detecting and Gauging Impact on Wikipedia Page Views. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1254–1261. https://doi.org/10.1145/3308560.3316751

[7] Wikipedia contributors. 2019. Kashmir Solidarity Day. https://en.m.wikipedia.org/wiki/Kashmir_Solidarity_Day [Online; accessed Oct 2019].

[8] Wikipedia contributors. 2019. Wikipedia:Emailing users — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Emailing_users [Online; accessed 12-Dec-2019].

[9] Wikipedia contributors. 2020. Administrators — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Administrators [Online; accessed May 2020].

[10] Wikipedia contributors. 2020. List of languages by total number of speakers — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers [Online; accessed May 2020].

[11] Wikimedia Contributors. 2020. Proposals for closing projects/Deletion of Bulgarian Wikinews. https://meta.wikimedia.org/wiki/Proposals_for_closing_projects/Deletion_of_Bulgarian_Wikinews [Online; accessed 11-Mar-2020].

[12] Wikipedia contributors. 2020. Single-purpose account — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Single-purpose_account [Online; accessed May 2020].

[13] Wikipedia contributors. 2020. Wikipedia:Core content policies — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies. [Online; accessed 14-Oct-2020].

[14] Wikipedia contributors. 2020. Wikipedia:Neutral point of view - Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view. [Online; accessed 14-Oct-2020].

[15] Wikipedia contributors. 2021. Languages of India — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Languages_of_India&oldid=1000337451 [Online; accessed on 15-Jan-2021; Page Version ID: 1000337451].

[16] Wikipedia contributors. 2021. Languages of Pakistan. https://en.wikipedia.org/w/index.php?title=Languages_of_Pakistan&oldid=999472017 [Online; accessed on 15-Jan-2021; Page Version ID: 999472017].

[17] dawn.com. 2019. *On Kashmir attack, Shah Mahmood Qureshi says 'violence is not the govt's policy.* Dawn.com. Retrieved December 2, 2019 from https://www.dawn.com/news/1464205

[18] N. Farič and H. W. Potts. 2014. Motivations for contributing to health-related articles on Wikipedia: an interview study. *J Med Internet Res* 16, 12 (2014), e260. https://doi.org/10.2196/jmir.3569

[19] Claudia Flores-Saviaga and S. Savage. 2019. Anti-Latinx Computational Propaganda in the United States. *ArXiv* abs/1906.10736 (2019).

[20] M Golebiewski and d boyd. 2018. Data voids: Where missing data can easily be exploited.

[21] Scott A. Hale. 2014. Multilinguals and Wikipedia Editing. In *Proceedings of the 2014 ACM Conference on Web Science* (Bloomington, Indiana, USA) *(WebSci '14)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/2615569.2615684

[22] Brent Hecht and Darren Gergle. 2009. Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies* (University Park, PA, USA) *(C&T '09)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/1556460.1556463

[23] Brent Hecht and Darren Gergle. 2010. *The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context.* Association for Computing Machinery, New York, NY, USA, 291–300. https://doi.org/10.1145/1753326.1753370

[24] Sydney Morning Herald. 2010. Two thirds in Kashmir want independence. https://www.smh.com.au/world/two-thirds-in-kashmir-want-independence-20100912-156b0.html.

[25] Joseph J Hobbs. 2008. *World regional geography.* Nelson Education, Boston, MA.

[26] Paul Jaccard. 1912. The Distribution Of The Flora In The Alpine Zone.1. *New Phytologist* 11, 2 (Feb 1912), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

[27] Brian C. Keegan. 2015. *Emergent Social Roles in Wikipedia's Breaking News Collaborations.* Springer International Publishing, Cham, 57–79. https://doi.org/10.1007/978-3-319-05467-4_4

[28] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/1240624.1240698

[29] Daniel Moyer, Samuel L Carson, Thayne Keegan Dye, Richard T Carson, and David Goldbaum. 2015. Determining the influence of Reddit posts on Wikipedia pageviews. In *Ninth International AAAI Conference on Web and Social Media.* AAAI Press Oxford, UK, AAAI Press, Oxford, UK, 75–82.

[30] Nida Najar. 2016. How Killing of Prominent Separatist Set Off Turmoil in Kashmir. https://www.nytimes.com/2016/07/17/world/asia/how-killing-of-prominent-separatist-set-off-turmoil-in-kashmir.html

[31] BBC News. 2019. *Kashmir: Why India and Pakistan fight over it.* BBC News. https://www.bbc.com/news/10537286

[32] BBC News. 2019. *Pulwama attack: India will 'completely isolate' Pakistan.* BBC News. Retrieved February 28, 2008 from https://www.bbc.com/news/world-asia-india-47249133

[33] BBC News. 2020. *Pulwama: Indian soldiers' families still waiting for justice.* BBC News. https://www.bbc.com/news/world-asia-india-51499752

[34] Government of India. 2019. *The Gazette of India : Extraordinary.* Government of India, New Delhi, India. http://egazette.nic.in/WriteReadData/2019/210049.pdf

[35] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '09)*. Association for Computing Machinery, New York, NY, USA, 51–60. https://doi.org/10.1145/1531674.1531682

[36] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '09)*. Association for Computing Machinery, New York, NY, USA, 51–60. https://doi.org/10.1145/1531674.1531682

[37] Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication* 12, 1 (2006), 88–113.

[38] Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic Vandalism Detection in Wikipedia. In *Advances in Information Retrieval*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 663–668.

[39] Reuters. 2007. Majority in Kashmir Valley want independence: poll. https://www.reuters.com/article/us-kashmir-poll/majority-in-kashmir-valley-want-independence-poll-idUSDEL29179620070813.

[40] Koen Smets, Bart Goethals, and Brigitte Verdonk. 2008. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. , 43–48 pages.

[41] Joan Soler-Adillon and Pere Freixa. 2017. Wikipedia access and contribution: Language choice in multilingual communities . A case study. *Anàlisi* 0, 57 (2017), 63–80. https://analisi.cat/article/view/n57-soler-adillon-freixa

[42] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 127 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359229

[43] Jacob Thebault-Spieker, Aaron Halfaker, Loren G. Terveen, and Brent Hecht. 2018. Distance and Attraction: Gravity Models for Geographic Content Production. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173722

[44] The Economic Times. 2019. Kashmir shuts down on Burhan Wani's death anniversary; Yatra, security convoy movement suspended - The Economic Times. https://economictimes.indiatimes.com/news/politics-and-nation/kashmir-shuts-down-on-burhan-wanis-death-anniversary-yatra-security-convoy-movement-suspended/articleshow/70131204.cms?from=mdr. "[Online; accessed on 12/12/2019]".

[45] Chau Tran, Kaylea Champion, Andrea Forte, Benjamin Mako Hill, and Rachel Greestadt. 2020 (forthcoming). Are anonymity-seekers just like everybody else? An analysis of contributions to Wikipedia from Tor.

[46] Marlon Twyman, Brian Keegan, and Aaron C. Shaw. 2016. Black Lives Matter in Wikipedia: Collaboration and Collective Memory around Online Social Movements.

[47] Upwork. 2020. Upwork.com. https://www.upwork.com/

[48] Wikipedia contributors. 2020. Translate us — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Translate_us [Online; accessed 1-Jun-2020].

[49] Wikipedia contributors. 2020. Translation — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Translation [Online; accessed 1-Jun-2020].

[50] Wikipedia contributors. 2020. With great power comes great responsibility — Wikipedia, The Free Encyclopedia". https://en.wikipedia.org/wiki/With_great_power_comes_great_responsibility [Online; accessed 15-Jan-2021].

[51] Dennis M. Wilkinson and Bernardo A. Huberman. 2007. Cooperation and Quality in Wikipedia. In *Proceedings of the 2007 International Symposium on Wikis* (Montreal, Quebec, Canada) *(WikiSym '07)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/1296951.1296968

[52] Xtools. 2019. XTools. https://xtools.wmflabs.org/. [Online; accessed 12-Dec-2019].

10 APPENDIX

A BACKGROUND: JAMMU AND KASHMIR

Ever since gaining independence from the British, Pakistan and India have been in conflict over Kashmir, a disputed region in the northern part of both countries, where they border each other. While both countries only control a part of the former princely state, both claim Jammu and Kashmir (J&K) in its entirety [25]. Three wars and several other conflicts have yielded the current line of control, which demarcates the regions administered by each of three nations: India, Pakistan, and China. An insurgency began to proliferate in India-administered Kashmir in the late 1980s.

B PULWAMA ATTACK OF 2019

Unrest in Kashmir grew in 2016 after Indian security forces killed a popular militant leader, Burhan Wani [30, 44]. On February 14, 2019, a convoy of vehicles carrying security personnel on the Jammu Srinagar National Highway was attacked by a vehicle-borne suicide bomber at Lethpora (near Awantipora) in the Pulwama district, Jammu and Kashmir. The attack resulted in the deaths of forty Central Reserve Police Force (CRPF) personnel and the attacker. The Pakistan-based Islamist militant group Jaish-e-Mohammed claimed responsibility for the attack. The attacker was Adil Ahmad Dar, a local of the Pulwama district and a member of Jaish-e-Mohammed [32]. India has blamed Pakistan for the attack. Pakistan condemned the attack and denied any connection to it [17].

C INTERVIEW QUESTIONS

C.1 Demographics

(1) Username
(2) Age
(3) Location - Country
(4) Languages (spoken at home, mother tongue, and others)
(5) Familiarity with English
(6) What is your occupation?
(7) How long have you been a Wikipedia user?
(8) How long have you been a wikipedia editor?
(9) Briefly in a few words describe why you use Wikipedia.

C.2 Goals & Practices

(1) Start by stating the article name(s) and language(s) (if interviewee edited more than one)
(2) Why do you edit Wikipedia articles?
(3) Do you ever refer to other language versions of Wikipedia? If so, why and how often?
    - Observations (inconsistencies, do you correct those?)
(4) How often do you find that you need to edit out vandalism
(5) How often do you find that you need to edit incorrect information
(6) How much do you interact with other editors?
    - How are your interactions?
    - How do you know if an editor is trustworthy?
(7) How much do you know about house rules and how well do you follow them? Have you ever enforced them?
(8) How often do you visit Talk pages?
(9) How much time do you spend on editing wikipedia articles?
(10) Regarding the following wikipedia article, why did you feel the need to edit this article?
(11) Would you consider yourself an expert on the topic?
(12) How do you source information and references for each edit?
    - How do you choose a hierarchy of importance?
    - How do you decide whether or not a source is biased or neutral?
(13) When and why did you feel compelled to edit this specific article?
(14) Did you edit articles related to this one?

(15) Have you edited Hindi or Urdu versions of this article? If so, did you use the same username?
(16) Do you ever refer to other language versions of this article and or related articles?
(17) What are your personal views on the topic that you have edited?

Received June 2020; revised October 2020; accepted December 2020