Wiki HUEs: Understanding Wikipedia practices through Hindi, Urdu, and English takes on evolving regional conflict

Molly G. Hickman, Viral Pasad, Harsh Sanghavi, Jacob Thebault-Spieker, Sang Won Lee
[mollygh,viralpasad,harshks,jthebaultspieker,sangwonlee]@vt.edu
Computer Science, Virginia Polytechnic Institute and State University
Blacksburg, Virginia

ABSTRACT

Wikipedia is the product of thousands of editors working collaboratively to provide free and up-to-date encyclopedic information to its 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. Additionally, we interviewed editors from all three Wikipedias to learn differences in editing processes and motivations. While activity on South Asian language editions of Wikipedia is growing, at the time of writing, the Hindi and Urdu editions are still in their nascency. In Hindi and Urdu, as well as English, editors predominantly adhere to the principle of *neutral point of view* (NPOV), and the editors quash attempts by other editors to push political agendas.

CCS CONCEPTS

• **Human-centered computing → Wikis**.

KEYWORDS

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 [16, 22]. The Wikimedia Foundation has fostered Wikipedias in over 250 languages, the most frequently edited and visited of them being the English edition [14]. The lack of information sharing between different language editions of Wikipedia prevents access to a larger variety of content for monolingual users [16].

The 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 [15]. Adversarial actors have been known to exploit Wikipedia pages and to manipulate search engine results to present a biased, false, or purposefully misleading version of events, thereby disseminating their propaganda to the broader internet [12, 13]. Therefore, it is significant for us to understand how wiki editors cope with the challenges in sustaining the neutral point of view (NPOV).

The goal of this paper is to understand how the editors of two lesser-edited 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. To this end, we conducted time series analyses of page view and revision data to learn how quickly articles are consumed and updated to reflect breaking news events. In addition, we interviewed editors to learn how their writing practices vary from what we know from previous works and how they maintain a neutral point of view (NPOV) on a topic that is prone to bias. We found many ways in which Hindi and Urdu editors differed from English editors in the way they collaborate, as well as one key similarity: editors in all three languages were devoted to presenting neutral accounts of the conflict.

2 JAMMU AND KASHMIR CONFLICTS

Jammu and Kashmir (J&K) is a union territory administered and claimed by India. Three-quarters of the region's population is Muslim. Both Pakistan (itself a Muslim majority Islamic Republic) and India claim the entirety of the Kashmir region. Three wars and various conflicts within the region have resulted in J&K being divided between India, Pakistan, and China. Thanks to Article 370 of the Constitution of India, J&K was granted autonomy in 1954. Hence, all laws passed by the Indian Parliament were not directly applicable to the state without the assent of the state's government. 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. Thereafter, the Home Minister of India introduced a reorganization bill in the Indian Parliament to divide the state into two union territories [6]. 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.

These factors motivated us to study the Kashmir conflict and its effects on Wikipedia. We chose to study the following Wikipedia articles:

- (A1) *Article 370 of the Constitution of India* (English, Hindi, Urdu)
- (A2) *Kashmir conflict* (English, Hindi, Urdu)
- (A3) *Insurgency in Jammu and Kashmir* (English, Hindi)

One of these three articles, titled *Insurgency in Jammu and Kashmir* on the English Wikipedia, has no counterpart on the Urdu Wikipedia.

3 RELATED WORK

Wikipedia has been actively researched, helping us to understand the participatory nature of collaborative, online knowledge sharing enabled by Web 2.0. Our study builds on previous research into the multilingual and multicultural nature of Wikipedia. Hecht et al. [16] 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 [15] highlights how self-focus in community-maintained knowledge repositories can be explained. Hale [14] has examined the roles of users editing multiple language editions of Wikipedia articles. He discusses how multilingual editors bring more context, sources, and perspective from their main language articles to articles in other languages, reducing the level of self-focus bias on Wikipedia. Pfeil et al. found several factors that differed between multiple cultures they studied, including reluctance to delete information, likelihood to collaborate with other editors, and propensity for spelling mistakes [20].

This work also draws ideas from previous works that studied how editors respond to breaking news events; Keegan has studied how Wikipedia editors collectively respond to breaking news events [19], and Keegan et al. studied the same phenomenon as it related specifically to the Black Lives Matter movement [21]. That work found that when the *Black Lives Matter* page is subject to a spike in attention and edits, often, other articles related to the Black Lives Matter movement are also viewed and edited.

4 DATA COLLECTION AND METHODS

Our focus is 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. Hence, we aim to answer the following research question: **How do Wikipedia practices of editors in the three languages vary in general and how they uphold the NPOV principle ?** We approached this question from the following perspectives.

- How quickly do editors in three languages respond to significant events of relevance to an article?
- To what extent do editors work on multiple articles across topics or languages?
- How do editing practices differ between languages (e.g., sources, collaboration, neutrality)?

We drew data on revisions for each article, from the time of creation through December 31, 2019, from the MediaWiki API[1]. The data, pulled on March 11, 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). We used R to pull daily page view data on each article, using the *WikipediR* package[2].

[1]https://www.mediawiki.org/wiki/API:Revisions
[2]https://cran.r-project.org/web/packages/WikipediR/index.html

4.1 Page views

We examined a page view time series for each article in order to validate our choice of articles and to 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 [19, 21]. The more significant an event is to a topic, the more views articles on that topic receive.

4.2 Revision Data

We also examined the sizes (in bytes) of the articles over time, which grow and shrink as editors add and remove content. Using the revision time series of each article, we can observe whether there exists a correlation between the edit spike times and volumes on the same article in English, Hindi, and Urdu.

To answer the question of how timely editors' responses are to significant events, we performed a time-lagged cross-correlation analysis on the views and edits of each article. Taking the two daily time series (views and edits), we computed the correlation between the views on one day and the edits on one day with an interval. The interval with the highest correlation indicates how many days editing activity lags behind viewing activity for each article in each language: a measure of how quickly articles "catch up" to current events.

4.3 Editor Pool Analysis and Interviews

To extract unique editors from Wikipedia, we used the XTools web tool[3] [5]. Editors of interest for the articles in our corpus were identified based on the sizes of their edits, time since their last edit, and total number of edits to the target articles. Each editor, barring anonymous editors, was sent a recruitment email using Wikipedia's *EmailUser* facility [4]. Editors were asked to participate in a one-hour, semi-structured interview [8].

Five editors were interviewed 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 extracted key elements and ideas discussed in the interviews using thematic analysis [10].

Age	Country	Primary Wikipedia	Language Proficiencies	Wiki Edits
23	IN	HI	HI, EN	<5K
22	IN	HI	Marwari, HI, EN	40–50K
40	PK	UR	UR, EN	20–30K
37	US	EN	EN, Spanish	2,000,000+
19	IN	EN	EN, HI, Bengali	10–20K

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

[3]https://xtools.wmflabs.org/

5 RESULTS

5.1 Editors Are Responsive to Significant Events

There were local maxima in page views per day on February 14, 2019, and August 5, 2019 (the dates of the Pulwama attack (see appendix B) and the revocation of Article 370, respectively) for all the articles in our corpus, with the possible exception of the two Urdu articles: neither saw as clear of a spike in views when the Pulwama attack occurred, compared to their counterparts in Hindi and English. There was a spike on February 5, 2019 on *Article 370* in Urdu. This can be attributed to Kashmir Solidarity Day, a national holiday in Pakistan, where Urdu is the national language [7]. Both *Article 370* and *Kashmir conflict* in Urdu did have a smaller swell in page views a few days after the Pulwama attack. While we were not actively searching for such a connection, the fact that people seeking information in Urdu on Kashmir Solidarity Day (February 5) referred to the two Urdu articles in our corpus further supports the idea that the articles we chose are of particular relevance to current events in J&K.

When events occur that are directly related to a given article, the rate at which editors respond varies by language edition. For example, 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 37,000 bytes in length for most of 2019, rose to over 60,000 bytes within days of the revocation. Meanwhile, the Hindi edition went from just under 18,000 bytes to almost 26,000, and the Urdu edition saw no significant change in article volume, staying at around 5,000 bytes. Figure 1 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.

The pages on *Article 370* were the only ones that showed a significant correlation between the volume of edits and page views over time. On the English article, edits and views are most highly correlated on the same day (r=0.84); that is, when page views are high on a given day, edits, too, are high on that day. Editors of the English *Article 370* page responded quickly to increased interest in the page; as we saw in Section 3.1, this tends to signify that an important event has just occurred. Edits to the Hindi edition are also moderately correlated (r=0.54) with same-day views, although the greatest peak correlation occurs with a five-day lag (r=0.59), meaning Hindi daily editing volume tends to spike either on the same day as page views spike, or five days from an event.

The data on the Urdu page on *Article 370* were inconclusive. There were only five days in 2019 when the article was edited, and only one of those days saw more than three edits. The most edits in a day in 2019 were made on February 7 (ten edits), two days after the aforementioned Solidarity Day and one week before the attack at Pulwama. With so little editing activity, any correlation between views and edits is likely to be due to noise.

English is a major language in the Indian subcontinent—as such, it is not surprising that English-speaking editors respond quickly to new developments on such topics as Article 370. Despite having

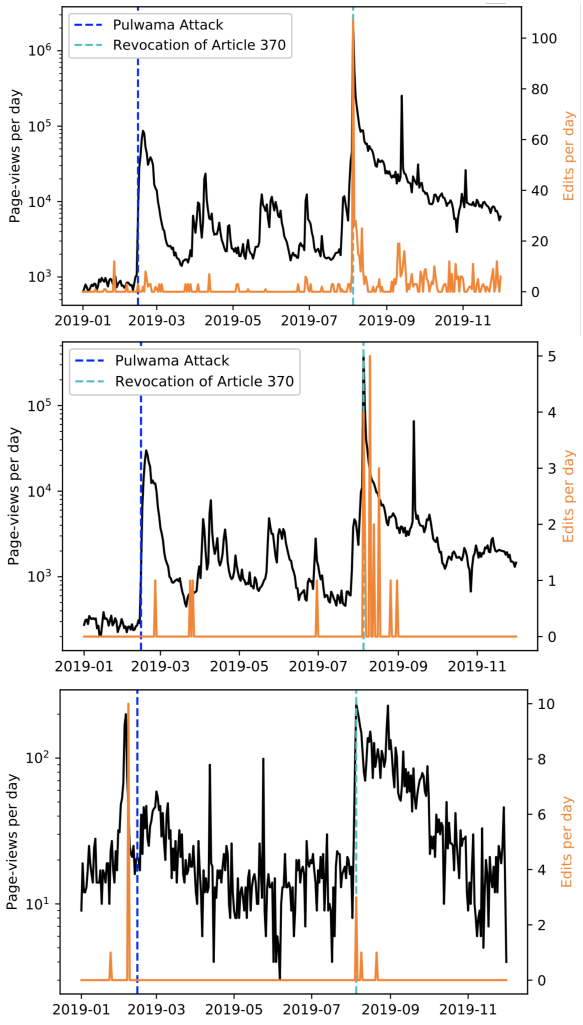

Figure 1: Edits and page views per day for *Article 370 (A1)* English (top), Hindi (middle), and Urdu (bottom)

far fewer editors, the Hindi page was also edited shortly after page views spiked.

5.2 Cross-Language Edits Were Infrequent

We examined the overlap between sets of editors of the two articles (A1, A2) that are available in all three languages. If a person edits a page without logging into an account, their IP address is displayed instead. We treated IP addresses and account names the same[4], to see whether users worked on multiple articles across two or three languages.

Even though English is a major secondary language in both countries (India and Pakistan), we rarely observed cross-language participation : the Hindi and English versions of A1 and A2 had one and three co-editors, respectively; the English and Urdu versions of A1 and A2 had two and no co-editors, respectively; and no editors edited both the Hindi and Urdu versions of either article. To compare editorship overlap, we computed an indicator of similarity

[4]Note that this is a proxy for the number of unique editors. One person may have multiple accounts or have made changes from multiple IP addresses.

between sets (the Jaccard index[5]), and saw essentially zero overlap in all three cases (on the order of 0.001). This does not imply that Indian and Pakistani editors do not tend to edit the English Wikipedia. Based on the locations of the IP addresses, we noticed that some of the edits of articles in English were made from India and Pakistan, and an interviewee confirmed that many only edit the English Wikipedia, it being the most widely read Wikipedia.

We examined if editors worked on both *Article 370 (A1)* and *Kashmir conflict (A2)* within the same language. We found that a similar, small proportion of Hindi editors (7 out of 86) edited both articles as did English editors (65 out of 2,015) (Jaccard index = 0.08 and 0.03, respectively). On the other hand, five out of 22 Urdu editors edited both articles (Jaccard index = 0.23).

5.3 Interview Results: Wiki Editing Practices

5.3.1 Motivations. All interviewees echoed a passion for and hobby of editing on Wikipedia. Editors of the Hindi and Urdu Wikipedias were interested in "growing the language" and "serving the readers of the language." One participant additionally listed freedom of access to free knowledge as a motivation. Overall, interviewees were interested in improving the quality and size of the articles.

5.3.2 Collaboration strategies. All of the Hindi and Urdu interviewees stressed the importance of communicating with their peers, while the two English editors interviewed preferred to work alone. The most visible collaborative interactions happened on the *article talk* or *user talk* pages. Interestingly, the Hindi and Urdu editors felt a stronger sense of community through remote and in-person collaboration rather than through the on-Wikipedia Talk pages. The Hindi and Urdu editors used various virtual and offline channels (e.g., Facebook, WhatsApp) to communicate with one another informally, delegate work, and discuss issues. The non-English interviewees also revealed that editors meet up regularly, especially if they are in close geographical proximity. The English editors interviewed were not actively part of any offline or virtual groups related to Wikipedia. We believe that the reason for this strong collaboration is the common goal of generating information on a relatively small edition of Wikipedia. Also, in the Hindi and Urdu editions, the editors were more invested in on-boarding and guiding newer editors on the platform than the English editors were.

5.3.3 Sourcing information. We expected to see self-focus bias or differences in the references chosen by editors of different languages on the same topic. Instead, we found that all the Hindi and Urdu editors used English Wikipedia articles as their primary sources, with some content being directly translated from there. Only if an editor personally knew that a local source was more correct than one in the English Wikipedia would they use it as a reference. The English editors, on the other hand, usually sourced information from traditionally "reliable" sources, such as reputable news organizations, public databases, and other sources, such as journals or scholarly articles.

5.3.4 Structure, Content, and Consistency. The English editors indicated that their focus on the editing process had moved towards maintaining the structure of the content, fixing the writing style of articles, and adding missing information, whereas the Hindi and Urdu editors focused on adding content; consequently, their pages were characterized by a relaxed approach to rules and structure.

5.3.5 Conflict, Propaganda, and Vandalism. Conflicts between editors happen because of differences in opinions on content or references, but often, also because of vandalism and propaganda. Anecdotal evidence showed that intra-language conflicts for the Hindi and Urdu Wikipedias were largely resolved through offline and online communication, while on-the-record (user and article talk page) conflicts were less common than on the English Wikipedia.

The English editors spoke about differences in opinion on structure, semantics, and references, although none lasted very long, largely being resolved through wars of attrition (editors who stand their ground on the pages for longer tend to get their way). All the interviewees provided anecdotal evidence of vandalism and propaganda by anonymous editors. For example, one interviewee spoke of a group of rogue editors spreading propaganda on the day of the revocation of Article 370. As another incident on the English *Kashmir conflict* page (A2), two editors made the same revision at two different times (January 22, 2019 and August 6, 2019), removing a section related to human rights violations by Indian forces against Kashmiris. On both occasions, the removal was undone by another editor within fewer than fifteen minutes [1, 2].

6 CONCLUSION AND FUTURE WORK

Despite having many fewer editors than the English Wikipedia, the Hindi and Urdu Wikipedias' pages in our corpus were edited quickly in response to breaking news events. The Hindi and Urdu editors we interviewed were invested in improving their respective editions of Wikipedia, upholding the NPOV principle, closely collaborating with one another, and using English Wikipedia pages as primary sources for material on their wikis. Vandalism, though prevalent, is ephemeral, thanks to rapid housekeeping by veteran editors in all three language editions. Attempts to weaponize these Wikipedia articles for political purposes seem to have failed quickly.

We plan to take a closer look at edits by anonymous users. Interviewees told us that most vandalism is committed by editors without usernames, and several speculated that some of these editors might be paid propagandists. Paid propaganda is a growing threat—not only to Wikipedia or the region we chose for this study, but to websites and organizations around the globe.

[5]The Jaccard index, or coefficient, here is the fraction of the cumulative unique editors of two articles who edited *both* articles; that is, the intersection of the two sets of editors divided by the union. The index ranges from 0 to 1, 1 being the most similar (all in common) and 0 being the most diverse (none in common) [18].

REFERENCES

[1] [n.d.]. Kashmir conflict: difference between revisions. https://en.wikipedia.org/w/index.php?title=Kashmir_conflict&type=revision&diff=909554720&oldid=909554572 (Accessed on 12/13/2019).

[2] [n.d.]. Kashmir conflict: difference between revisions. https://en.wikipedia.org/w/index.php?title=Kashmir_conflict&type=revision&diff=879703886&oldid=879700840 (Accessed on 3/11/2020).

[3] [n.d.]. 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. (Accessed on 12/12/2019).

[4] [n.d.]. Wikipedia:Emailing users - Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Emailing_users. (Accessed on 12/12/2019).

[5] [n.d.]. XTools. https://xtools.wmflabs.org/. (Accessed on 12/12/2019).

[6] 2019. *The Gazette of India : Extraordinary.* Government of India. http://egazette.nic.in/WriteReadData/2019/210049.pdf
[7] 2019. Kashmir Solidarity Day. https://en.m.wikipedia.org/wiki/Kashmir_Solidarity_Day
[8] 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.
[9] bbc com. 2019. *Pulwama attack: India will 'completely isolate' Pakistan.* Retrieved February 28, 2008 from https://www.bbc.com/news/world-asia-india-47249133
[10] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).
[11] dawn.com. 2019. *On Kashmir attack, Shah Mahmood Qureshi says 'violence is not the govt's policy.* Retrieved December 2, 2019 from https://www.dawn.com/news/1464205
[12] Claudia Flores-Saviaga and Saiph Savage. 2019. Anti-Latinx Computational Propaganda in the United States. *CoRR* abs/1906.10736 (2019). arXiv:1906.10736 http://arxiv.org/abs/1906.10736
[13] M Golebiewski and D Boyd. 2018. Data voids: Where missing data can easily be exploited. *Data & Society* 29 (2018).
[14] Scott A. Hale. 2014. Multilinguals and Wikipedia Editing. In *Proceedings of the 2014 ACM Conference on Web Science* (Bloomington, Indiana, USA) *(WebSci '14)*. ACM, New York, NY, USA, 99–108. https://doi.org/10.1145/2615569.2615684
[15] 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)*. ACM, New York, NY, USA, 11–20. https://doi.org/10.1145/1556460.1556463
[16] Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-generated Content and Its Applications in a Multilingual Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 291–300. https://doi.org/10.1145/1753326.1753370
[17] Joseph J Hobbs. 2008. *World regional geography.* Nelson Education.
[18] 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
[19] Brian Keegan. 2014. *Emergent Social Roles in Wikipedia's Breaking News Collaborations.* Springer.
[20] 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.
[21] Brian C.; Shaw Aaron Twyman, Marlon; Keegan. 2016. Black Lives Matter in Wikipedia: Collaboration and Collective Memory around Online Social Movements. (2016).
[22] 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

7 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 [17]. 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

According to Time, unrest in Kashmir grew in 2016 after India killed a popular militant leader, Burhan Wani [3]. 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, India. The attack resulted in the deaths of forty Central Reserve Police Force (CRPF) personnel and the attacker. Responsibility for the attack was claimed by the Pakistan-based Islamist militant group Jaish-e-Mohammed. The attacker was Adil Ahmad Dar, a local of the Pulwama district and a member of Jaish-e-Mohammed [9]. India has blamed Pakistan for the attack. Pakistan condemned the attack and denied any connection to it.[11]