Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model

ELIJAH MAYFIELD, Language Technologies Institute, Carnegie Mellon University, USA
ALAN W BLACK, Language Technologies Institute, Carnegie Mellon University, USA

In this work we show that machine learning with natural language processing can accurately forecast the outcomes of group decision-making in online discussions. Specifically, we study *Articles for Deletion*, a Wikipedia forum for determining which content should be included on the site. Applying this model, we replicate several findings from prior work on the factors that predict debate outcomes; we then extend this prior work and present new avenues for study, particularly in the use of policy citation during discussion. Alongside these findings, we introduce a structured corpus and source code for analyzing over 400,000 deletion debates spanning Wikipedia's history, enabling future large-scale studies of group decision-making discourse.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Machine learning*; • **Information systems** → World Wide Web.

Additional Key Words and Phrases: group decision-making; Wikipedia; machine learning; online communities

ACM Reference Format:
Elijah Mayfield and Alan W Black. 2019. Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 206 (November 2019), 26 pages. https://doi.org/10.1145/3359308

1 INTRODUCTION

In group decision-making tasks, members participate in a constrained discussion, where they must choose from a fixed set of possible outcomes and there is no objective right answer. Participants debate the merits of the different options; correctness is often a judgment call with persuasive arguments in favor of more than one outcome [43, 51]. In these tasks, dysfunction leads to poor performance, with low-quality discussion that fails to effectively fit together information from different group members [62]. High-performing groups by contrast have consistent characteristics like shared values, mental models of the problem, and communication styles, resulting in nuanced patterns of conflict and consensus-building [57]. Group decision-making extends to online settings, where users in online production communities want to make good choices that will improve their collaboration over time. They accomplish this through intricate systems of social norms and cues for resolving disputes [35]. But the details of how these decisions are made can be difficult to analyze or measure quantitatively. While much of online decision-making happens in free-form text discussions, quantitative research often ignores the rhetorical and discursive details of this practice, *"observing change from before to after the deliberation without considering what has happened during the discussion"* [58]. Detailed analysis of discourse practices in online decision-making has seen less

Authors' addresses: Elijah Mayfield, elijah@cmu.edu, Language Technologies Institute, Carnegie Mellon University, USA; Alan W Black, awb@cs.cmu.edu, Language Technologies Institute, Carnegie Mellon University, USA.

2573-0142/2019/11-ART206 $15.00
https://doi.org/10.1145/3359308

research activity compared to study of easier-to-quantify metadata (like number of posts, social network degree statistics, or other aggregated measures) [18].

Here, we demonstrate one way that researchers can benefit from metrics of group interactions that go beyond raw counts. In our recent work [42], we introduced a new model for forecasting outcomes of online group decision-making discussions. This allows us to produce a task-relevant, normalized measures of the shift in likely outcomes that can be associated with each observed contribution in a discussion. We argue that this forecasting model can isolate interesting and relevant discourse phenomena correlated with debate outcomes, a valuable tool for social scientists in their analysis of groups. In domains like collaborative learning, this approach has already been validated for analysis [47, 52] and for development of interventions like automated conversational agents [1]. Here we show that a similar principle can benefit the broader study of group decision-making online, analyzing a particular domain with a long history of research: Wikipedia deletion debates. Alongside publication of this work, we release a public, preprocessed corpus of hundreds of thousands of debates from Wikipedia, with timestamped votes and comments, and outcomes. Our work lays the foundation for future research in three ways:

- We show that a forecast model for group decision-making outcomes can differentiate editor behaviors that are hard to distinguish with more straightforward methods. This greater detail allows us to learn new things about Wikipedia discussion norms.
- By releasing our corpus of Wikipedia debates, we invite future work on that discourse community, either extending our research questions or opening new avenues of inquiry.
- By demonstrating our methods in the Wikipedia context, we set the stage for broader contributions to group decision-making research on sites outside Wikipedia and in offline settings.

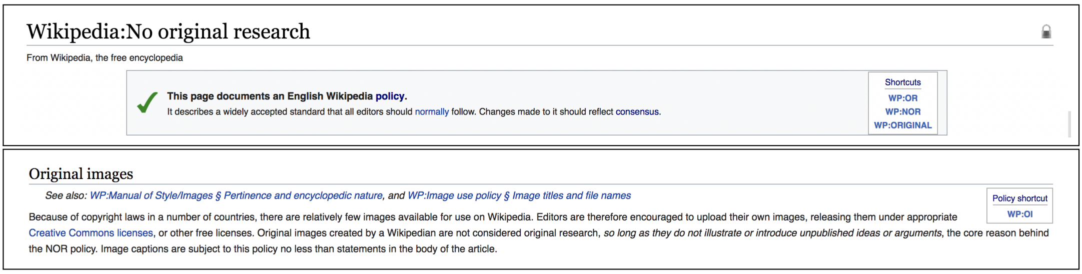
In the first portion of this paper (section 2), we review the historical context of administrative tasks that led to Wikipedia's current decision-making infrastructure. We introduce our corpus in section 3, giving comprehensive statistics about this domain for the first time since [60] in 2010, and introduce our method for forecasting group outcomes using machine learning in section 4. Next, in section 5 we produce a series of results that this forecast model enables, with a special focus on how *policy citation* can be analyzed to better understand the dynamics of Wikipedia editorial discussions. Based on these findings, we take time in section 6 to recommend an agenda for future work, both within Wikipedia and in group decision-making more broadly.

2 BACKGROUND ON WIKIPEDIA

Countless papers have studied Wikipedia (see Mesgari et al. [45] for a thorough survey), and a subset have studied editor interactions as a "model organism" for decision-making online[1]. Though Wikipedia was first founded in 2001, it took a few years for research interest in the community of editors to begin in earnest. The earliest published research on Wikipedia was likely [41]; shortly thereafter, numerous articles appeared in the following year[2]. Much of this early work focused on editor motivation and hierarchy formation, trying to determine how high-quality writing could reliably emerge from spontaneous editor communities [11]; this early work built a foundational understanding of editor incentives and stratification that still informs much of today's research [15, 63]. The growth in traffic caused a shift to maintenance work and internal debate rather than creation of new content [38]. This radical refocusing from content authoring to bureaucracy, led to a *"rational effort to organize"* through policies and guidelines [5]. While our work focuses on deletion debates, prior work has also studied deliberation and argument on in other administrative venues, like *Requests for Comment* [29] and on talk pages [2].

[1] This term, used in the context of social media analysis, originates with Tufekci [61].

[2] Of these, the comparison of Wikipedia's accuracy to *Encyclopaedia Brittanica* in [21] was most widely disseminated.



Wikipedia:No original research

From Wikipedia, the free encyclopedia

✓ **This page documents an English Wikipedia policy.**
It describes a widely accepted standard that all editors should normally follow. Changes made to it should reflect consensus.

Shortcuts
WP:OR
WP:NOR
WP:ORIGINAL

Original images

See also: WP:Manual of Style/Images § Pertinence and encyclopedic nature, and WP:Image use policy § Image titles and file names

Policy shortcut
WP:OI

Because of copyright laws in a number of countries, there are relatively few images available for use on Wikipedia. Editors are therefore encouraged to upload their own images, releasing them under appropriate Creative Commons licenses, or other free licenses. Original images created by a Wikipedian are not considered original research, *so long as they do not illustrate or introduce unpublished ideas or arguments*, the core reason behind the NOR policy. Image captions are subject to this policy no less than statements in the body of the article.

Fig. 1. Top: Header of the `No original research` policy, which can be linked using aliases (`OR`, `NOR`, and `ORIGINAL`). Bottom: one specific subsection of that policy, which can be linked directly (`WP:OI`).

After this expansion period, the site experienced a long, steady decline in the following decade. The slowdown was noted almost immediately [59] and attributed to three factors: an increase in overhead necessary for "maintenance" and administrative tasks for the larger community; newcomers turning away due to exclusion and gatekeeping from existing editors; and structural resistance to new edits through page protection and reverts. The pattern of slowing activity continued for several more years as the site matured and newcomer participation became even more difficult [23]. While early decisions were made "by fiat" from user leaders or site founders, this was replaced over time by a decentralized network of committees, administrators, policies, and decision-making forums [17]. Wikipedia's norms for editor interaction are "highly conservative" and long-lived in comparison to most other online communities [35]. Today, much authority on the site remains grounded in a small network of policies shaped early in the site's history [8], written originally in response to a period of heavy growth and necessary crowd control. Some of the earliest policies, like `Notability` (`N`), `Verifiability` (`V`), and `No Original Research` (`NOR`, see Figure 1) originated many years ago but continue to dominate editorial discussion and drive group decision-making, while newer rules remain comparatively obscure [27].

This effect is partially due to newcomer behavior, which has trended away from spending time editing the site's core content in favor of time spent in discussion on talk pages, administrative disputes, and bureaucracy [35]. This activity has value — for instance, citing policies on article talk pages *does* influence editing behavior [48] — but ingrained norms and expectations can have an effect of preventing new or underserved users from meaningfully contributing to Wikipedia [36]. In particular, Wikipedia's editor culture has a highly visible and discussed gender gap[3]. An early study found that among the 3% of site editors who self-reported gender, only 9% of edits are made by women [40]. Further studies showed this gender gap exists among high-skill internet users only; in low-skill or unsophisticated web users, no gender gap was observed [26]. Later work found that in social production communities, rigid policies and norms as well as complex user interfaces *reinforce* a gender gap over time rather than reducing disparities [14]. While a stream of new users is a necessary part of the site's continued ability to thrive, basic site functionality like edit reverts has had a significant, ongoing demoralizing effect on newcomers [24].

Over the last several years, Wikipedia has worked to bring newcomers into their community, but predicting what will be effective is challenging in such a complex domain. But while some approaches to improving retention have worked well [46], other attempts have backfired and been shown to *decrease* productivity of new users [54]. Not all contributions are created equally; recent work has begun to recognize the granular intentions of individual edits to more intelligently

[3] Note though that all cited studies on gender gaps in Wikipedia assume a biological and binary gender. This is a methodological exclusion of non-binary and transgender individuals common in HCI research; for more, see Keyes [37].

categorize editor actions [68, 69], and this approach has been useful for understanding users. For instance, emotional labor roles that are necessary to maintain basic community functioning have lower associated prestige, and those roles are disproportionately filled by women editors [44]. Yet work has shown that raw edit counts and basic forms of politeness are sufficient predictors of administrator promotion [4], and users seeking promotion can “pad their stats" with relatively minor edits and other easily measurable editorial actions [9]. Building on this research tradition, our work aims to design and validate more granular methods for differentiating types of user contribution and the part that editors play in decision-making processes.

2.1 *Articles for Deletion*

We analyze Wikipedia’s *Articles for Deletion* discussion domain. Editors at *AfD* nominate pages to these discussions when they believe they should be removed from the wiki, and usually include a nominating statement giving a rationale for deletion. After nomination, a discussion is held open for at least seven days[4]. When a page is nominated to *AfD*, any user (including unregistered users, provided they sign their post with an IP address) can place a vote, which must include a rationale for why they believe an article should be kept or removed from the wiki. These votes are public, signed, and timestamped. Users can also make non-voting comments, either in direct reply to the nomination, or in reply to a vote or other comments, in the standard “reply tree" model of online discussion forums [3]. *AfD* is highly active, with more than one third of *all* articles in the English-language administrative namespace `Wikipedia:` related to deletion debates.

Discussions are aggregated by an administrator, who determines the discussion outcome. This is not a popular vote; the final tally of a debate is not the deciding factor, though administrators rarely deviate from majority votes. Administrators may also hold debates open for a longer period of time, or close discussions with a verdict of `No consensus`. If no consensus is reached, nominated articles are kept by default; deleting articles requires an unambiguous outcome.

As with the rest of Wikipedia, *AfD* is subject to a broad set of written and unwritten norms for social behavior. Many of these norms have been encoded into hundreds of written and highly visible policies, guidelines, or essays[5]. A long-running ideological divide in these debates exists on a spectrum between “deletionist" and “inclusionist." The former stance prefers high standards for material, culling less broadly relevant content and emulating the historical role of encyclopedias as gatekeepers. The latter stance argues for a reshaped role of information sources online, including, at its most extreme, *any* potentially valuable information that can be independently verified.

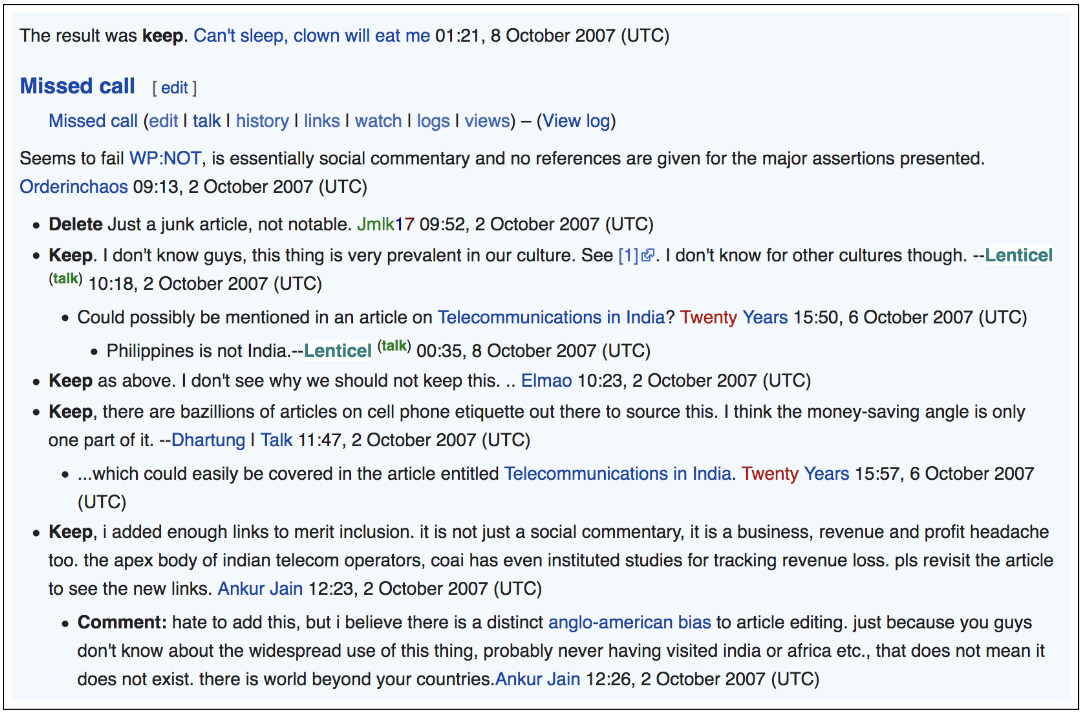
Figure 2 gives an example of these dynamics in practice, for the article *“Missed Call."* As part of the opening nomination for deletion, the nominating editor cites the `Wikipedia is not a dictionary` policy (abbreviated `WP:NOT`) and lack of sources:

> *“Seems to fail* `WP:NOT`*, is essentially social commentary and no references are given for the major assertions presented."*

User preferences - mostly for `Delete` and `Keep`, with a long tail of alternate options - are highlighted in **bold**, with some users voting to remove the page, and some to keep. Going back to our example, the nomination is followed up by votes with rationales:

> “**`Delete.`** *Just a junk article, not notable."*
> “**`Keep.`** *I added enough links to merit inclusion. It is not just a social commentary, it is a business, revenue and profit headache too.* [. . .] *pls revisit the article to see the new links."*

[4] Policies also allow “speedy" resolution, skipping this timeline for exceptions like libel or plagiarism of copyrighted material.
[5] These are terms of art, clearly denoted by page templates. Policies reflect a mandatory consensus, guidelines contain generally accepted principles, and essays give advice without broad acceptance. For more detail, see Forte & Bruckman [16].



The result was **keep**. Can't sleep, clown will eat me 01:21, 8 October 2007 (UTC)

Missed call [edit]

Missed call (edit | talk | history | links | watch | logs | views) – (View log)

Seems to fail WP:NOT, is essentially social commentary and no references are given for the major assertions presented. Orderinchaos 09:13, 2 October 2007 (UTC)

- **Delete** Just a junk article, not notable. Jmlk17 09:52, 2 October 2007 (UTC)
- **Keep**. I don't know guys, this thing is very prevalent in our culture. See [1]. I don't know for other cultures though. --Lenticel (talk) 10:18, 2 October 2007 (UTC)
  - Could possibly be mentioned in an article on Telecommunications in India? Twenty Years 15:50, 6 October 2007 (UTC)
    - Philippines is not India.--Lenticel (talk) 00:35, 8 October 2007 (UTC)
- **Keep** as above. I don't see why we should not keep this. .. Elmao 10:23, 2 October 2007 (UTC)
- **Keep**, there are bazillions of articles on cell phone etiquette out there to source this. I think the money-saving angle is only one part of it. --Dhartung | Talk 11:47, 2 October 2007 (UTC)
  - ...which could easily be covered in the article entitled Telecommunications in India. Twenty Years 15:57, 6 October 2007 (UTC)
- **Keep**, i added enough links to merit inclusion. it is not just a social commentary, it is a business, revenue and profit headache too. the apex body of indian telecom operators, coai has even instituted studies for tracking revenue loss. pls revisit the article to see the new links. Ankur Jain 12:23, 2 October 2007 (UTC)
  - **Comment:** hate to add this, but i believe there is a distinct anglo-american bias to article editing. just because you guys don't know about the widespread use of this thing, probably never having visited india or africa etc., that does not mean it does not exist. there is world beyond your countries.Ankur Jain 12:26, 2 October 2007 (UTC)

Fig. 2. Excerpt from a single AfD discussion, with a nominating statement, five votes, and four comments displayed. Keywords labeled in "'**bold**"' are explicit preferences from users, which are treated as votes.

Discussion consists of followup comments, as well as action: as shown in the example above, users may take proactive measures to *improve* an article during an *AfD* nomination, in accordance with the principles of an open, user-generated Wiki. In this example, after eight votes and thirteen comments from ten total participants, an administrator closed the discussion with a Keep outcome.

.

2.2 Prior work on debate in Wikipedia *AfD*

Substantial work on *AfD* has already taken place. The first detailed study of deletion decisions took place in Taraborelli and Ciampaglia [60]. This work found a herding effect among participants, where later votes were highly influenced by the early tally of votes. It also found that user voting patterns could be well-described with a clustering model that contained only two centroids, coarsely corresponding to "inclusionist" and "deletionist" users. The findings suggested retained preferences of individual users over time, and made recommendations for more sophisticated analyses to come.

Next, a comprehensive early study attempted to directly quantify the *quality* of *AfD* debates [39]. They approached this problem by looking for articles that were deleted but later re-created, or kept but later re-nominated for deletion. They found a number of factors that led to good decision quality, like larger group size, groups that were diverse in experience level (but not groups heavy on recruited users or newcomers), and decisions made by unbiased administrators.

Geiger and Ford [20] later analyzed debates and found a deep disconnect between the participants in debates and the authors that produced content. In particular, they found that an overwhelming majority of debates included no first-time participants at all, and that it was rare for article authors

Table 1. Summary of key findings from prior *AfD* studies. Our released corpus of 423k debates 2005-2018 contains a superset of all data in these papers, except early debates from 2003-04 in [60].

Prior Work	**Corpus**	**Key Findings**
Taraborelli & Ciampaglia [60]	223k debates 2003-10	Early voters cause "herding." Individual users maintain `Delete`/`Keep` preferences across debates.
Lam et al. [39]	158k debates 2005-09	Larger groups with a diversity of tenure produces better decisions. Recruiting creates biased groups but does not hurt decision quality. Bias of individual administrators can lower quality.
Geiger & Ford [20]	120k debates 2007-11	Small groups dominate *AfD*. Article creators rarely participate. 96% of participation comes from repeat editors and 74% of debates have no newcomers.
Joyce et al. [33]	588 debates pre-2012	Vote tallies and comment activity predict outcomes. Admin influence on outcomes is not significant. Citing the `WP:IAR` policy helps `Keep` votes.
Schneider et al. [55, 56]	72 debates, Jan. 2011	Novices and experts use different arguments. Both can be ineffective: novices make ineffective use of policy, while experts lean too much on boilerplate.
Xiao et al. [30, 64–67]	Subsets from 2010-2015 (229, 5k, 39k debates)	Notability dominates *AfD* rationales. Some topics, like biographies, have more unanimous outcomes than others. `Keep` votes have more positive sentiment. Expert editors frequently give imperative commands to newcomers.

to participate in the discussion about their own article (under 20% of discussions). Later, Schneider et al. performed a qualitative review of 72 debates [55], conducting a close read of specific debates and gave additional observations on the divide between readers and editors, the obscure requirements and norms placed upon newcomers, and ordering effects that prioritized early votes in debate outcomes. Following this work, Joyce et al. [33] tested a series of hypotheses on how rules and hierarchies interact with success in *AfD*. They replicated the prior finding that votes do predict outcomes, but added nuance on the application of seniority and policy, making several observations and doing a close study of two policy categories in particular, `Notability` and `Ignore All Rules`. The former policy was found to be universally predictive of successful votes, while the latter was correlated with success for `Keep` votes, but not `Delete`.

More recently beginning in 2014 with [65], Xiao et al. undertook a series of mixed-methods studies of rationales in *AfD* votes. They again replicated the finding that vote counts did significantly predict outcomes. Contradicting Joyce et al., they found that `Notability` topics were still the most common topic of argument in the domain but found no significant correlation between outcomes and the percentage of notability citations in discussions. They also surveyed *topics* for likely outcomes, and found significant relationships: biographies and for-profit companies were more likely to be deleted than other topics, while locations and events were more likely to be kept. Later work by the same researchers has moved to discourse analysis in the *AfD* domain, such as the use of sentiment analysis [67], imperatives [66], and tree-style data visualization [30], without making outcome prediction a central research question.

We situate our investigation in this body of work, with key results summarized in Table 1; these results will be referred to throughout our analysis and replicated when possible.

Table 2. Overall breakdowns of labels across all data.

	Delete	Keep	Merge	Redirect	Other
Votes (2005-2018)	54.9	28.4	3.6	3.8	9.3
Outcomes (2005-2018)	63.9	20.7	3.2	6.0	6.2
Prior Work [60] (2003-2010)	63.6	23.6	3.9	1.9	7.0

3 CORPUS DESCRIPTION AND BASELINE ANALYSIS

We created a large offline, preprocessed corpus of *Articles for Deletion* discussions. This snapshot contains the full text of all *AfD* debates in the English-language Wikipedia from January 1, 2005 to December 31, 2018. Prior to 2005, community norms, discussion formatting, and deletion process were more erratic, making automated extraction difficult and limiting any findings even if the data was successfully extracted[6]. In addition to the raw text, this corpus is structured with extracted metadata, specifically timestamps, outcomes, nominations, votes, users, and policy citation. A total of 402,440 discussions were extracted. For analysis, we then filtered out two categories of discussions, mostly from earlier years when formatting norms were less standardized:

- Discussions without an outcome label from an administrator (20,669 instances, or 5.1%).
- Discussions that received no votes after nomination (12,179 instances, or 3.0%).

After these exclusions, our analysis covers 369,592 debates. Our corpus contains a more comprehensive set of debates than any prior work, both more recent and more thorough, nearly doubling the raw size of the largest existing studies. Due to this, replication or failure to replicate results is not an admonition of prior work, and may be a product of sample size and time period rather than a contradictory finding.

Table 2 shows percentages for each label for vote and outcome distributions in the analyzed subset. To analyze policy norms, we manually assembled a list of frequently cited links in *AfD* discussions. Editors can link to overall policy pages or directly to subsections; additionally, many pages and subpages can be linked using any of a number of shortcut aliases. The taxonomy we built includes 37 policy pages with 377 sections, 44 guideline pages with 398 sections, and 71 essay pages with 201 sections, all linked by a total of 2,111 shorthand aliases. For each contribution, we extract all hyperlinks to any alias in our taxonomy. While this is a reasonable proxy for policy citation, it is not comprehensive of all use of policy, for at least three reasons:

- Most citations are added intentionally by the editor who signs the contribution; however, some are added after the fact (like links to the `SIGNATURES` policy, appended by bots to unsigned posts, along with the username or IP address logged for the contribution).
- While this taxonomy includes all official policies and the vast majority of guideline and essay citations in *AfD*, there is a long tail of rarely-cited essays and pages that are not comprehensively included in our taxonomy and were not extracted.
- We do not capture citations to policies without MediaWiki links to those pages (merely writing "NBIO" to refer to the notability policy on biographies, for instance, instead of writing "[[WP:NBIO]]" to include a link). Editors generally follow formatting conventions and include links when appropriate, but this is a source of missing data.

The preprocessed data for our analysis will be released in parallel with the publication of this paper. This includes the full corpus, including the 8.1% of filtered nominations with no discussion

[6] For this same reason, the corpus study in Lam et al. [39] also chose the January 2005 starting point.

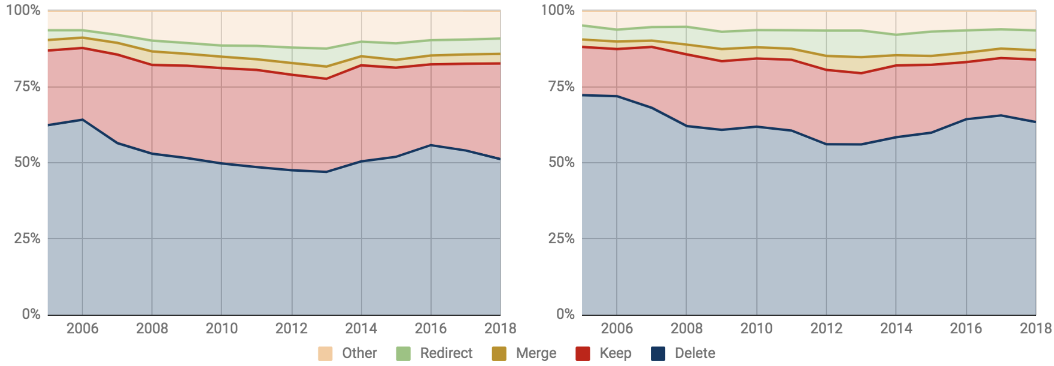


Fig. 3. Distributions by year for votes (left) and outcomes (right) over Wikipedia's history.

or missing outcomes; labels for all votes and outcomes, in three levels of granularity[7]; and the manually constructed taxonomy of policies, guidelines, essays, and aliases. Appendix A includes a sample of the format used for this data in JSON format. Additionally, we hope that the NLP and machine learning methods presented here are generalizable for behavioral science research on group decision-making, particularly in online settings. To facilitate this, all source code available is released under an open source license[8].

3.1 Vote and Outcome Distributions

Our corpus makes available the first comprehensive review of activity statistics in *AfD* since 2010 [60]. Figure 3 shows distributions of voting preferences over time, separating votes and outcomes. Vote totals approximately match reported distributions from work early in Wikipedia's history; however, we find a much narrower spread between Delete and Keep votes compared to early work. While that work showed a 40-point margin in favor of Delete (64% to 24%) [60], we only observe a 26.5-point margin. Additionally, we measure distributions of final administrative *outcomes*, and find that outcomes are more deletionist than votes, with Keep comprising over 28% of votes but fewer than 21% of final outcomes. Part of this is driven by the increased length and controversy of discussions that lead to Keep outcomes - more votes are cast per debate than in uncontroversial Delete decisions. Additionally, the presence of long-tail rare labels is more common in outcomes than in votes, such as outcomes with multiple actions (Merge and Delete, for instance), or outcomes resulting in No Consensus (which defaults to a Keep outcome, functionally).

This gap is partially explained by the difference in time period observed in our dataset. Delete votes were already becoming less common in the later years of that study's window of observation, a pattern that has since been maintained. The decline in site activity was linked to a continued decrease in Delete votes, falling from a peak of 64.1% in 2006 to a low of 47.0% in 2013, then seeing a modest resurgence but mostly stabilizing over the last decade at levels lower than the early peak.

Mirroring the overall drop in editor activity over time, voting activity in debates has declined over time. After reaching a peak of 6.9 votes per discussion in 2006, activity declined, and discussions have averaged 4.3 votes in the past ten years. Figure 4 gives volume counts for discussions over

[7] The released corpus can be normalized to a two-label model using only Keep and Delete, or to a five-vote model that also maintains separate categories for Merge, Redirect, and Other. We also preserve raw text of votes and outcomes, which includes a very long tail of free-form inputs. For all analyses in this work, we use the two-label model, but we release source code for using the five-label variant.

[8] https://github.com/emayfield/AFD_Decision_Corpus

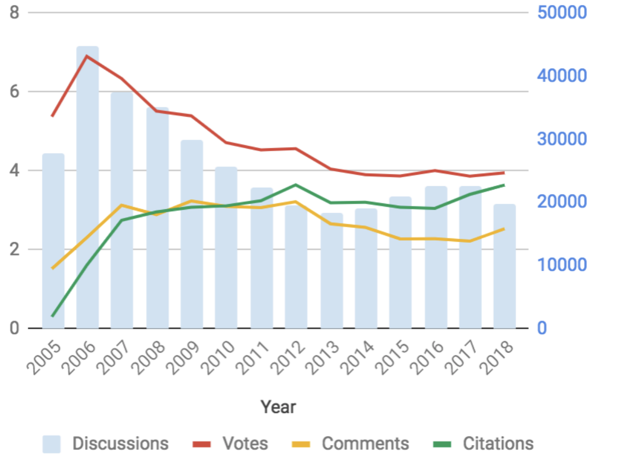


Fig. 4. Counts of discussions per year (blue) and of votes, comments, and citations *per discussion* in each year.

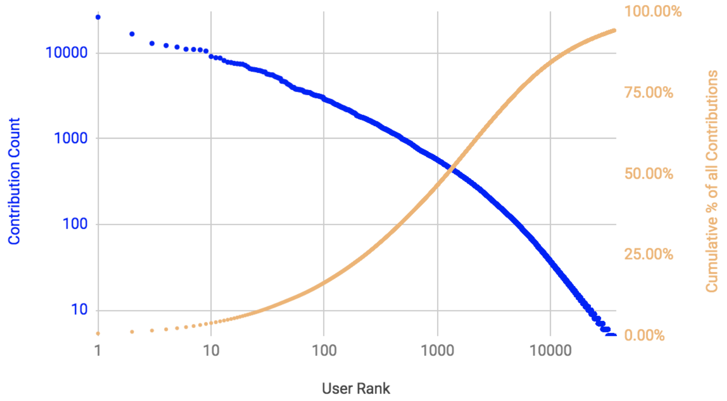


Fig. 5. Log-log plot of user rank and contributions. The top 36,440 users, all with at least five contributions, are displayed. Collectively, these 22.6% of all users account for 94.3% of all contributions.

time. In contrast to the decline in voting activity, we see a slow and steady growth of citation to policy. In early years of the site, votes outnumbered citations to policy by a ratio of more than ten to one. In the most recent year, vote and citation counts are near parity.

We also measure success rates for votes, defining a user's vote as successful if its label matches a keyword in the outcome decided by an administrator. Across all votes in AfD's history, 67.9% have been successful; this number rises to 75.6% when only considering votes for `Keep` or `Delete` outcomes and excluding votes for rare outcomes. Overall, deletionism is more successful: `Delete` votes are successful 82.0% of the time, while `Keep` votes have a 64.0% success rate.

The full set of contributors to our corpus is made up of over 161,266 editors in a log-normal distribution[9], visualized in Figure 5. Half of all contributions are made by 1,218 users, or just under 0.8% of editors present in our corpus. In contrast, 124,826 observed users (77.4%) contributed fewer

[9] By log-likelihood ratio, log-normal more closely fits contribution counts than other heavy-tailed distributions, $p < 0.01$.

than 5 edits; cumulatively, they account for only 5.7% of the observed data. Most frequently, users enter *AfD* to participate in a single debate, and never return. These results replicate the observation from prior work [20, 55] that *AfD* is dominated by long-time members rather than newcomers; in fact, as this trend has increased in recent years, the distributions we observe are *more* extreme than what has been previously reported.

4 A FORECASTING MODEL FOR GROUP DECISION-MAKING

One natural task that we immediately investigated upon collecting this corpus was our ability to predict outcomes using natural language processing and machine learning, with the text of discussions as input[10]. That work detailed the technical aspects of prediction; the work presented here, in contrast, uses the *output* of that model to give greater insight into the underlying discussions, and test whether those findings align with prior work on *AfD*. But digging into the model in such a way requires at least an overview of the machine learning that drives the analysis, and so we review here our past experimental setup and model performance. We use the following notation:

- A single deletion discussion is labeled d. It has a series of contributions $[c_0, c_1 \ldots c_N]$.
- Each contribution c_i has a corresponding username u_i, vote label l_i (null for comments), timestamp t_i, and a rationale text, r_i (which might be empty).
- The features of a single contribution c_i can be extracted using arbitrary representations of language, and represented as ϕ_i.

We use standard natural language processing features to represent the text of vote and comment rationales ϕ_i. Traditionally, language has been represented as a "bag-of-words," the standard representation of text data for decades, still in widespread use [34]. More recent representations of text have used word *embeddings*, representing language not as a single feature but as dense vectors pre-trained from large unsupervised corpora. An example of this style of model is $GloVe$ [50], which we evaluate. The newest embedding models are *contextual*: rather than encoding word semantics as a fixed vector, words are represented based on their surrounding context at classification time. The most effective contextual model to date, the $BERT_{BASE}$ model [10], produces 768-dimensional embeddings for text sequences of up to 512 consecutive words, and enables state-of-the-art accuracy on numerous classification tasks. *BERT* was already trained on Wikipedia texts (and other sources), so we perform no fine-tuning[11].

We encode overall *discussion* content at a given timestamp t_i as $\phi_d(t_i)$, the average vector of each rationale text in contributions that have appeared up to that point, normalized by the length of each rationale's text in raw tokens:

$$\phi_d(t_i) = \frac{\sum_{j=0}^{i} \frac{\phi_j}{\ln(len(r_j))}}{i}$$

Representing entire discussions ϕ_d is then the case where all contributions are included, $\phi_d(t_N)$. For our machine learning prediction, we train a logistic regression classifier implemented in Scikit-Learn [49] with L2 regularization and the LIBLINEAR solver [13]. After this model has been trained, for a new input discussion d, the classifier predicts a probability distribution over discussion outcomes, $P(l|\phi_d)$, where probabilities sum to 1. The trained model is a forecast, given the data observable at a moment in time, of the likely outcome of a debate given the expressed preferences of a group.

[10] That work has been previously published; citation removed for peer review.

[11] This may mean text from our corpus is included in $BERT_{BASE}$ training data, causing a minuscule exposure to test data in our experimental setup; we do not investigate this question here.

Representation	**Full Debate**		**Incremental**	
	%	κ	%	κ
Majority Class Baseline	74.0	0.00	62.1	0.00
GloVe	81.7	0.49	69.1	0.31
Bag-of-Words	84.2	0.58	72.4	0.39
BERT	85.8	0.62	73.4	0.41
BERT + Vote Labels	93.5	0.83	79.7	0.55

Table 3. Forecasting accuracy for full discussions and incremental predictions from prior work [42]).

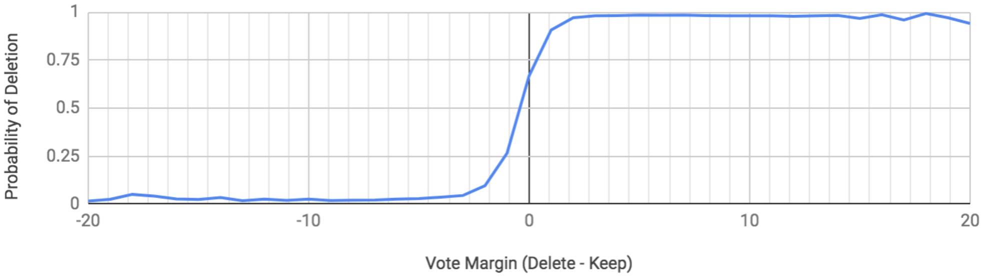


Fig. 6. Probability of a Delete outcome as voting margin varies. Administrators almost never overrule Delete majorities with a margin of at least 2 votes, or Keep majorities with a margin of at least 4 votes.

4.1 Forecasting Accuracy

We estimate accuracy of this model using 10-fold cross-validation. All instances from a particular discussion appear in only one fold; there is never crossover from the same debate between train and test data. We evaluated accuracy of this model on a randomized subset of 5% of the corpus, approximately 20,000 discussions. Table 3 reprints results that have been reported in our prior work. The *BERT* model reaches the best levels of agreement, outperforming other representations of text by at least 1.6% accuracy, in absolute terms. Short discussions are more predictable, with the best-performing model reaching accuracy of 97.3% for short discussions of 5 or fewer total contributions that resulted in a Delete outcome, compared to 85.3% accuracy for long discussions of more than 10 contributions that resulted in a Keep outcome.

We can then append additional features to our representation ϕ_d. For each possible vote label, we extract features including the raw count of votes for each label up to time t_i, and percent of votes at that point that have been cast for each option. This addition improves all models, with the *BERT* model still performing best, with Cohen's $\kappa = 0.83$. Models that use *only* vote tallies are highly accurate; in fact, the forecast model that takes language into account does not differ significantly in accuracy compared to a model using only gold labels. Only 7.6% of votes end in ties (administrators choose Delete in 66.9% of these cases), and as shown in Figure 6, outside of ties administrators follow the majority vote in 94.8% of discussions. Therefore the gold labels are highly informative features. Nevertheless, the content in contribution text encodes information that is vital for analysis, and so we include both text embeddings and vote label features in our model moving forward.

We next test our ability to make incremental predictions at earlier moments during discussions. We train our models identically in this set of experiments, using full discussions as training instances. The resulting output model is identical in both cases; only evaluation differs. In our test set we create a new instance for classification after *each* contribution to each discussion. Note that reported accuracy in this setup overweights more contentious debates - with more contributions, there are more instances from that discussion to classify in each test set. This slight bias results in over-representation of debates that ended in a Keep outcome, as those debates tend to have more contributions, and therefore increases the difficulty of the problem (Keep is a minority label and more challenging to predict). In this evaluation, all models see significant performance degradation, with lower accuracy from forecasting early in the debate. *GloVe* and bag-of-words models are more competitive, but *BERT* maintains the highest accuracy, with an overall accuracy of 79.7% across all instances and $\kappa = 0.55$, when vote labels are included.

4.2 Measuring Forecast Shifts

For the remainder of this work, we use the model that is given access to all observable information at training time, including both the gold labels and text of individual contributions. Using this model, we measure shifts in probability output from our forecast model at each of these incremental predictions. We measure the change in the posterior probability distribution of outcomes immediately after each contribution is posted[12]:

$$\Delta(l, c_i) = P(l|\phi_i) - P(l|\phi_{i-1})$$

This approach follows practices from prior work on “disparate impact," [7] which measures difference in expected outcomes given circumstances that differed by exactly one variable (in our case, time). Increase in forecasted probability of one label shifts that label upward, and another simultaneously downward, doubling the cumulative impact of changes; therefore, we sum the change in probability of outcomes for all labels and introduce a normalizing factor of ½ to produce a measure of *forecast shift*, ranging from [0,1] for each contribution.

$$\text{Forecast Shift} = \frac{1}{2} \sum_{l \in L} |\Delta(l, c_i)|$$

For corpus study, we measure forecast shift of contributions through 10-fold cross-validation. For each fold, we build a model on the training set, then make incremental probability forecasts using that model for each discussion in the test set. Iterated across each fold, we are able to measure forecast shifts for our entire corpus, with no discussion received forecasts from a model where that discussion was itself part of the training set.

4.3 Limitations

Our measure of forecast shift is a descriptive measure of how a predictive model alters its prediction based on new evidence. However, we have only inspected what is *predictive* given limited information, not what is rhetorically *influential* to the debates themselves. Thus while the features identified as shifting forecasts are informative, we cannot claim they are causal. As a crucial example of where this limits our analysis, we do not include article texts themselves in our work. Our predictive model has found that debate-initial Keep votes are predictive of Keep outcomes. This has, at minimum, two possible explanations. The first is that early Keep outcomes are persuasive or influential in the debate itself and lead to articles being preserved. The second is that articles worth preserving *attract* early Keep votes, and the rhetorical strategy of the voter is unimportant compared to the voter as proxy measure of article quality. We discuss alternative approaches for future work that could help discern causality in section 6.

[12] For nominations ($i = 0$), for each possible outcome $l \in L$, we instead subtract the overall prior probability distribution $P(l)$ as measured from training data.

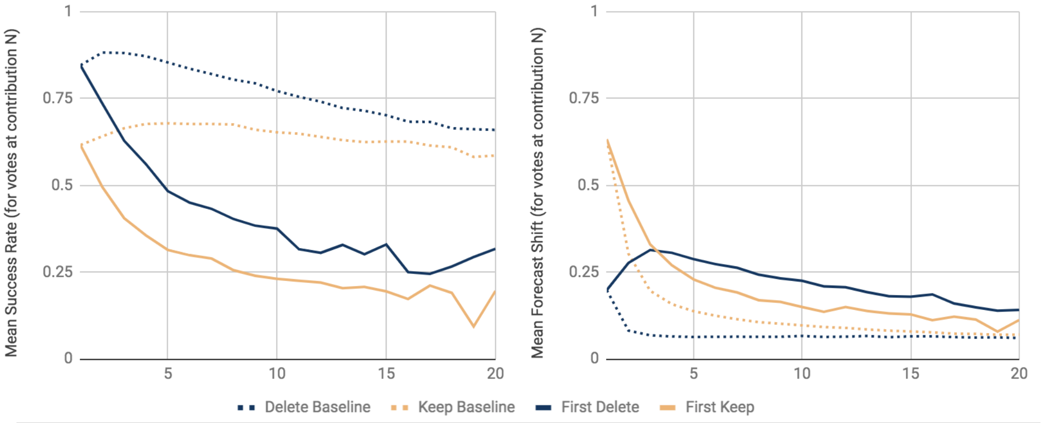


Fig. 7. Success rates (left) and forecast shifts (right) for votes that were the Nth contribution to a discussion, for different values of N. We measure these values first for *any* vote with that label at that ordinal location in the debate, then for discussions where the *first* vote for a particular label appeared at rank N.

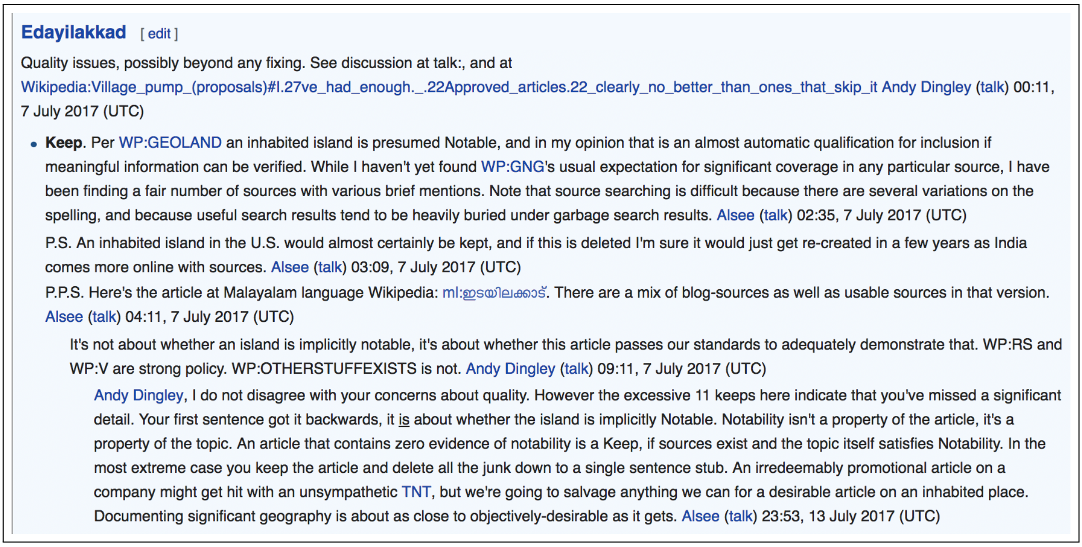
This is a common limitation in studies of *AfD*. Similar limitations applied to the early analysis in [60] that showed a “herding" effect, where later votes in a discussion tended to follow early votes; this too may have been because votes are an accurate proxy of article quality rather than a rhetorical impact of early voters on those who participate later. Notably, the studies that have focused on *rhetoric* in debates have avoided tying these analyses to success rates. For instance, in [67], researchers measured sentiment of votes and found more positive affect in Keep votes, but did not correlate sentiment to outcomes; similarly, in [56], researchers investigated the rhetorical argument strategies of editors in *AfD* but did not measure how those strategies affected success rates or influenced future decisions.

5 ANALYSIS WITH THE FORECAST MODEL

Within each of the analyses to follow, we will begin each section with a reference to the key prior work that informs a particular question. Along with measuring any particular user behavior, like arriving early or posting frequently, we will also single out specific policies that are often cited in exemplar cases of that behavior. This is a useful analytic lens; policies have consistently been a focus area of *AfD* research, from the close study of the `Ignore All Rules` policy in [33] to the study of notability subpolicies in [65]. The broad theme of our findings is that the relationship between policy citation, success, and forecast shift is nuanced. Many successful policies do not tend to appear in contributions that changed our model’s forecasted outcome, and many contributions that changed our model’s predicted outcome — sometimes dramatically — do not end up on the winning side of debates. We find that there is no overall correlation between the success rate of votes in which a policy has appeared, and the mean forecast shift from those votes; in fact the slope is very slightly negative ($r = -0.04$). Rather than describing the effect of policy citation as a monolithic phenomenon, we must instead study specific policies and how they are cited in context.

5.1 Early Voters

We first evaluate whether early contributions are correlated with discussion outcomes, following on observation of a “herding" effect in [60]. In that work, the authors found that later votes in discussions were more likely to mirror early votes. Our results replicate this finding: early votes



Edayilakkad [edit]

Quality issues, possibly beyond any fixing. See discussion at talk:, and at Wikipedia:Village_pump_(proposals)#I.27ve_had_enough._.22Approved_articles.22_clearly_no_better_than_ones_that_skip_it Andy Dingley (talk) 00:11, 7 July 2017 (UTC)

- **Keep**. Per WP:GEOLAND an inhabited island is presumed Notable, and in my opinion that is an almost automatic qualification for inclusion if meaningful information can be verified. While I haven't yet found WP:GNG's usual expectation for significant coverage in any particular source, I have been finding a fair number of sources with various brief mentions. Note that source searching is difficult because there are several variations on the spelling, and because useful search results tend to be heavily buried under garbage search results. Alsee (talk) 02:35, 7 July 2017 (UTC)

  P.S. An inhabited island in the U.S. would almost certainly be kept, and if this is deleted I'm sure it would just get re-created in a few years as India comes more online with sources. Alsee (talk) 03:09, 7 July 2017 (UTC)

  P.P.S. Here's the article at Malayalam language Wikipedia: ml:ഇടയിലക്കാട്. There are a mix of blog-sources as well as usable sources in that version. Alsee (talk) 04:11, 7 July 2017 (UTC)

  It's not about whether an island is implicitly notable, it's about whether this article passes our standards to adequately demonstrate that. WP:RS and WP:V are strong policy. WP:OTHERSTUFFEXISTS is not. Andy Dingley (talk) 09:11, 7 July 2017 (UTC)

  Andy Dingley, I do not disagree with your concerns about quality. However the excessive 11 keeps here indicate that you've missed a significant detail. Your first sentence got it backwards, it is about whether the island is implicitly Notable. Notability isn't a property of the article, it's a property of the topic. An article that contains zero evidence of notability is a Keep, if sources exist and the topic itself satisfies Notability. In the most extreme case you keep the article and delete all the junk down to a single sentence stub. An irredeemably promotional article on a company might get hit with an unsympathetic TNT, but we're going to salvage anything we can for a desirable article on an inhabited place. Documenting significant geography is about as close to objectively-desirable as it gets. Alsee (talk) 23:53, 13 July 2017 (UTC)

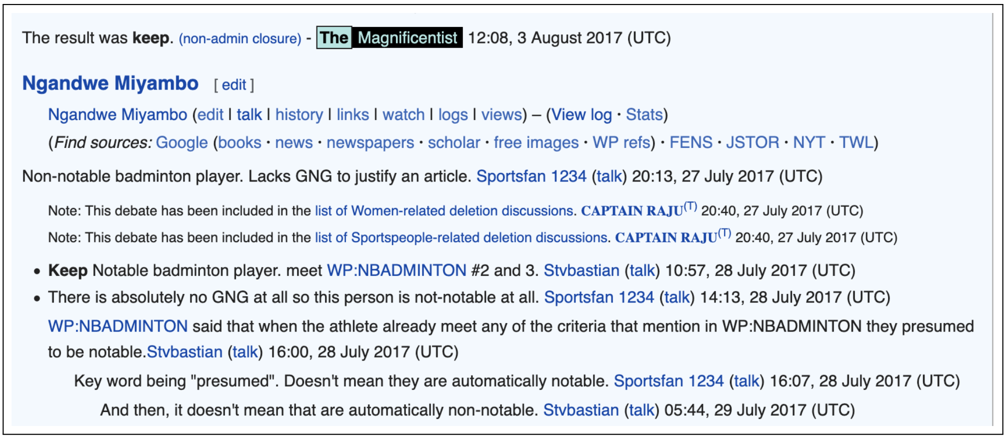
Fig. 8. Large forecast shifts arise from initial votes for Keep followed by response votes for Delete. Here, a user successfully cites the Notability (geographic features) policy to keep an article.

are highly predictive of outcomes. Debate-initial Delete votes are successful 84.5% of the time compared to a 63.9% baseline. The effect is even greater for early Keep votes, resulting in a Keep outcome 62.2% of the time, compared to a 20.7% baseline.

Trends over the course of a discussion are visualized in Figure 7. We separate the values for *all* votes that appear as the Nth contribution to a discussion from contributions at that point that were the first vote for a particular outcome. Success rates rapidly decline for both Delete and Keep when they arrive late in a discussion. When measuring forecast shift of votes, we see similar declines for votes overall, regardless of whether they are for Delete or Keep. Where we see differentiation is in forecast shift associated with the first Keep and Delete vote in a discussion. Early Keep votes are highly informative for the forecast model, and produce the greatest shift in forecast probabilities. But for Delete votes, arriving slightly later to a discussion *increases* forecast shift, peaking at the third contribution and only declining slowly when the first Delete vote appears later than that.

These results are intuitive. The default outcome when discussions do not reach consensus is Keep; however, the momentum in *AfD* is toward deletion. For inclusionist voters, a key factor in highly predictive votes is simply being *early* to arrive in a debate, either shifting the tenor of the discussion that follows or signaling clear article quality or meeting criteria for inclusion. When voting Delete, on the other hand, forecasts do not shift when they arrive early; a Delete outcome was already likely. Instead, Delete voters shift forecasts when they arrive in the middle of conversations and *contradict* earlier votes. The Delete voter shifts forecasts more significantly when acting as a "devil's advocate" and *reducing* certainty of a particular outcome; this is not possible in debates where deletion is obvious, and those Delete votes result in low values of forecast shift.

An example of this pattern is shown in Figure 8, where one user defend a page for a sparsely populated island in the Indian state of Kerala. We omit the (lengthy) discussion from this figure, but to summarize: the first voter produces a large forecast shift, beginning the debate with an initial Keep vote only two hours after nomination. When later users argue for Delete, the model shifts back to predicting a Delete outcome, but with low certainty. The next non-voting comment



The result was **keep**. (non-admin closure) - The Magnificentist 12:08, 3 August 2017 (UTC)

Ngandwe Miyambo [edit]

Ngandwe Miyambo (edit | talk | history | links | watch | logs | views) – (View log · Stats)

(*Find sources:* Google (books · news · newspapers · scholar · free images · WP refs) · FENS · JSTOR · NYT · TWL)

Non-notable badminton player. Lacks GNG to justify an article. Sportsfan 1234 (talk) 20:13, 27 July 2017 (UTC)

Note: This debate has been included in the list of Women-related deletion discussions. CAPTAIN RAJU(T) 20:40, 27 July 2017 (UTC)

Note: This debate has been included in the list of Sportspeople-related deletion discussions. CAPTAIN RAJU(T) 20:40, 27 July 2017 (UTC)

- **Keep** Notable badminton player. meet WP:NBADMINTON #2 and 3. Stvbastian (talk) 10:57, 28 July 2017 (UTC)
- There is absolutely no GNG at all so this person is not-notable at all. Sportsfan 1234 (talk) 14:13, 28 July 2017 (UTC)

WP:NBADMINTON said that when the athlete already meet any of the criteria that mention in WP:NBADMINTON they presumed to be notable.Stvbastian (talk) 16:00, 28 July 2017 (UTC)

Key word being "presumed". Doesn't mean they are automatically notable. Sportsfan 1234 (talk) 16:07, 28 July 2017 (UTC)

And then, it doesn't mean that are automatically non-notable. Stvbastian (talk) 05:44, 29 July 2017 (UTC)

Fig. 9. Highly successful votes that also shift our forecast model often come from the narrow use of established policies for notability in specific subtopics.

from the initial Keep voter gives detailed responses and further citation to policies, which tilts the forecast toward an eventual Keep outcome.

5.2 Notability Policies

A differentiating feature of the voter in the previous example is the citation of relevant and targeted policy, Notability (geographic features). In prior work, this citation behavior was previously noted as highly relevant for further study [65]; in particular, they noted that locations, biographies, and corporate pages were deleted at substantially different rates compared to pages in general. Our research extends that finding: Notability policies are among the most informative votes in our forecast model, appear early in debates (particularly often in Keep votes), and are more successful in general than other policies and more than votes in general.

Successful policy citations that *also* have high forecast shift are narrowly scoped. The most successful inclusionist Notability policies are on topics like astronomical objects, geographic landmarks, and local high schools. Enthusiasts wrote these policies to clearly define notability for an area where the average editor may not know inclusion criteria, and cite these policies effectively, first to shift the focus of discussions and then to win those debates. Some communities, though, reverse this trend and maintain highly selective standards to prevent an influx of articles; this phenomenon is most prevalent in sports, with highly successful citations in favor of Delete for topics like regional football (soccer) leagues and martial arts. In either case, forecast shift remains extremely high relative to all other policies, even as success rates differ dramatically. For a prototypical example of highly successful citations, we highlight the actions in Figure 9. This user is one of the top five most consistent users in our corpus, measured both by average success (over 90%) and average forecast shift of their posts. When they contribute to discussions, their votes are early in discussions and include clear citations to relevant policy in otherwise short rationales. By referencing criteria in a pre-existing policy (Notability (Badminton)), debate is closed quickly.

But as Notability policies become broader, their trends in both success rates and forecast shifts revert to the broader mean of all votes that cite policies. The very broadly scoped policy on proposed deletion of biographies of living people (WP:BLPPROD) is noteworthy: among all policies we study, it has the greatest difference in success and forecast shift metrics depending on whether it appears in Delete or Keep votes. When used in inclusionist arguments, the policy is usually

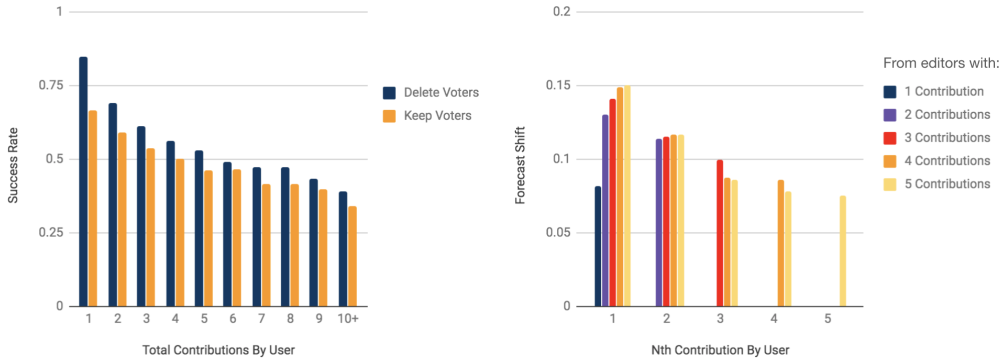


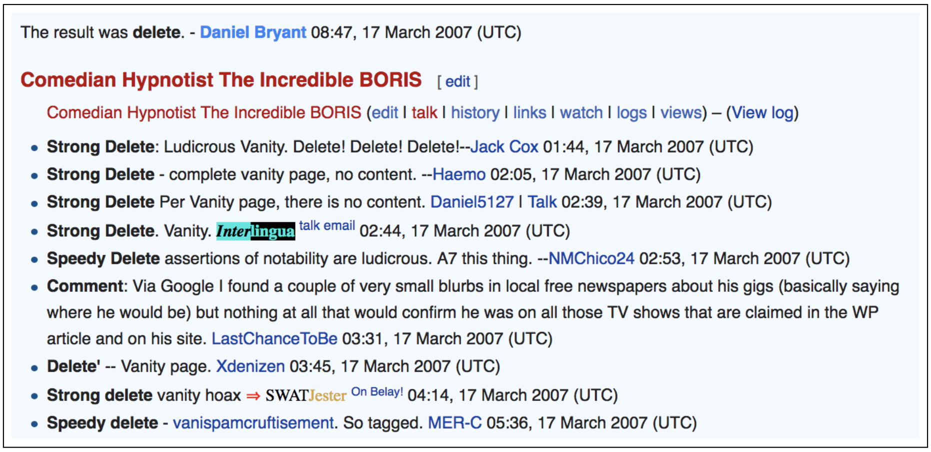
Fig. 10. One-time voters are more successful than more active voters; however, the first contribution from more active voters have greater forecast shift than the votes from one-time contributors.

cited early in the discussion and causes significant uncertainty in the model, shifting probable outcomes from being weighted toward delete to a tossup. However, those Keep citations of the biography policy are among the least successful votes in our corpus. By contrast, when cited as part of Delete arguments, this broad policy does much less to shift forecast probabilities, but is successful well above the baseline success rate for deletionist votes. Another way of seeing the role of Notability policies in debate is to look at the Delete policy citations with high average forecast shifts. While Keep votes have disproportionately high forecast shift values, the top two Delete citations as measured by average forecast shift are Trivial Mentions and Existence ≠ Notability. Both of these policies are used as *responses* to notability arguments from Keep voters.

5.3 Active Voters

Next, we measure whether votes from more *frequent* posters, who take the time to reply to other users and participate actively in discussion, are more predictive of future outcomes. This effect has been previously suggested in small-scale, mixed-methods analyses of dozens or hundreds of discussions [33, 65]. In our larger-scale corpus, though, we find mixed support for these findings. In our data, 45.0% of votes and comments in discussions come from editors that made more than one contribution in that discussion. Of these, single-contribution voters are substantially more likely to cast a successful vote, winning 84.8% of Delete votes and 66.7% of Keep votes. Users who post more than twice to a discussion are successful in fewer than half of their votes, and success rates continue to decline as users post more and more. This seems to contradict the topline finding from past work. As mentioned previously, this result is not causal: we cannot discern whether editors with weak arguments tend to add more comments to discussions, or whether their heavy participation in debates is in fact part of what leads administrators to side *against* those users in debates. But in either case, we do not find evidence that active users are more likely to win debates.

We find that relative to success rates as a measure, the forecast shift metric is actually a closer match for the observations from prior work. As shown in Figure 10, while success rates go down as users are more active in debates, the average forecast shift attributable to the first vote from those users is much higher. Forecast shifts are greater for the first post by editors who will eventually follow up with more activity; the first post by these highly active users (the lighter-shaded bars) shifts forecasts by almost twice as much as the first post by one-time contributors (the dark leftmost line). Additional contributions from those users, though, have diminishing returns.



The result was **delete**. - Daniel Bryant 08:47, 17 March 2007 (UTC)

Comedian Hypnotist The Incredible BORIS [edit]

Comedian Hypnotist The Incredible BORIS (edit | talk | history | links | watch | logs | views) – (View log)

- **Strong Delete**: Ludicrous Vanity. Delete! Delete! Delete!--Jack Cox 01:44, 17 March 2007 (UTC)
- **Strong Delete** - complete vanity page, no content. --Haemo 02:05, 17 March 2007 (UTC)
- **Strong Delete** Per Vanity page, there is no content. Daniel5127 | Talk 02:39, 17 March 2007 (UTC)
- **Strong Delete**. Vanity. Interlingua talk email 02:44, 17 March 2007 (UTC)
- **Speedy Delete** assertions of notability are ludicrous. A7 this thing. --NMChico24 02:53, 17 March 2007 (UTC)
- **Comment**: Via Google I found a couple of very small blurbs in local free newspapers about his gigs (basically saying where he would be) but nothing at all that would confirm he was on all those TV shows that are claimed in the WP article and on his site. LastChanceToBe 03:31, 17 March 2007 (UTC)
- **Delete'** -- Vanity page. Xdenizen 03:45, 17 March 2007 (UTC)
- **Strong delete** vanity hoax ⇒ SWATJester On Belay! 04:14, 17 March 2007 (UTC)
- **Speedy delete** - vanispamcruftisement. So tagged. MER-C 05:36, 17 March 2007 (UTC)

Fig. 11. Example of highly successful editor behavior with minimal forecast shift. For each of the later votes, the probability of a `Delete` outcome is already well over 99%.

As an example of this dynamic, we highlight the debate activity in Figure 11, arguing about a Canadian magician. In this debate, several users are successful in their vote, but do not meaningfully contribute to the decision-making process; in our forecast model, only one vote shifts the predicted outcome by more than 0.05, the very first by `Jack Cox`. By the time votes appear from later users, the discussion is a foregone conclusion for `Delete`. The late citation of `Vanispamcruftisement`, a lighthearted anti-spam policy, has no bearing on the clear consensus of the group. While single-vote users are highly successful, they are not changing the outcome of debates; instead, those late arrivals are getting credit for participation in a debate that has essentially concluded. Recognizing such behavior has high value for downstream applications, like decision support systems for administrator nominations. Other citations to policies about spam and hoaxes follow a similar pattern: they are among the top policies in our dataset when ranked by success rate, but consistently appear in votes with very little new information for our forecasting model.

5.4 Discussion Breakdowns

The previous finding that late arrivers have high success rates and yet do little to shift forecasts from our model; we can replicate that finding not with users but with policy by singling out the `Snowball Clause` policy, summarized as: *“If an issue does not have a snowball’s chance in hell of being accepted by a certain process, there’s no need to run it through the entire process."* This policy is cited once it is clear that consensus has been reached and that there is no need to hold discussion open for the full seven days. Indeed, we find that votes citing this policy have the highest success rate and *lowest* forecast shift of any policy in our taxonomy. Citing the `Snowball` policy in `Keep` votes is similarly at the bottom of our list of policies sorted by forecast shift.

We can also examine *other* policies that appear very late in discussions. We sort policies by the mean ordinal rank of the post in which they appear; in Table 4, we present the top-ranked policies on each end of this measure[13]. These policies are procedural and often indicate a breakdown in debate, with little information for our model to shift the likely outcome of debate. Instead, they are indicators that the debate’s content-focused discussion has ended, an outcome is highly likely, and debate decorum has now broken down entirely. This includes citations to policies like `No`

[13] For clarity: as shown in Table 4, `WP:Civility` citations appear in the 26th contribution to a discussion, on average.

Table 4. Policies sorted by the ordinal rank of when they appear in discussion, and the mean forecast shift of votes where that citation appears, split by vote label. Many early-appearing policies overlap with the influential notability policies from Table 4.

	Avg.	Forecast Shift			Avg.	Forecast Shift	
Earliest Citations	**Rank**	**Keep**	**Delete**	**Latest Citations**	**Rank**	Keep	Delete
Living Person Biographies	2.7	0.31	0.12	Civility	25.9	0.11	0.06
"Garage Bands"	3.3	N/A	0.09	No Personal Attacks	24.8	0.10	0.06
Notability (Media)	3.3	0.25	0.06	Attack Pages	23.6	N/A	0.07
Notability (Astronomy)	3.4	0.29	0.12	Disruptive Editing	22.6	0.12	0.04
Notability (Martial Arts)	4.0	0.27	0.09	Gaming the System	21.1	0.09	0.06
Notability (Music)	4.1	0.21	0.08	Arguments to Avoid	20.4	0.13	0.08
No Hoaxes	4.6	0.16	0.05	Ignore All Rules	19.4	0.13	0.07

personal attacks, Gaming the System, and No legal threats. Voters that cite these policies are on the losing side of debates, posting very late, and also are not changing the direction of those debates in which they appear. These results also tie into previous work from [33]. Studying a randomly selected set of 588 debates, those authors focused in on the Ignore All Rules policy, and suggested it had a significant effect. That policy has been cited a total of 1, 361 times in our corpus, usually *very* late in discussion, appearing on average in the 19th contribution. Among the votes in which the policy was cited, Keep votes were successful 10.3% less often than in the corpus overall, and appearing in successful Keep votes 53.7% of the time compared to a 64.0% baseline, and Delete votes dropped in success rates by 12.2%. Less contentious but also prevalent is procedural citation to editing guidelines, such as Editing Policy and Article Size. These votes, typically used in debates about lists that have been separated out of main articles and into separate standalone pages, tend to come very late in discussions.

6 DISCUSSION

A scatter plot showing the full distribution of policies analyzed for this study appears in Figure 12. We separate policies by their appearance in Delete and Keep votes. Policies that we highlighted earlier in our analysis are labeled. This is a busy figure and we take the time below to analyze its component pieces in detail. Overall, we find that because Delete votes are more successful, so too are citations that appear in those votes, but that observing any one of these votes does not tend to produce a large shift in probable outcomes in our forecast model. As a result, policy citations from Delete votes cluster in the top left of our scatter plot. Citations in Keep votes cause a much greater shift in our forecast model, as seen by the nearly clean partitioning of blue and orange clusters in our scatter plot. These policy citations are not necessarily successful, but do make the final outcome far less certain for our forecast model.

This builds an intuition for what our forecast shift metric is measuring. In our analysis of notability, we found that forecast shift for Delete voters in particular increased when citing policies like Trivial Mentions in response to Keep voters. In our analysis of activity, we found that forecast shift was associated with highly active users in debates even when those users were not successful. In the context of prolonged discussions, forecast shift might best be used to measure uncertainty or dissension, with more granularity than using post counts alone.

Our forecast model (and the metrics we derive from it) is a useful analytic tool for studying discussions; as we showed here, the forecast shift measure aligns neatly with findings from past work. Additionally, the measure does not require any explicit labeling of preferences or votes at the granularity of turns or even individuals. This makes the metric well-suited to other discussion

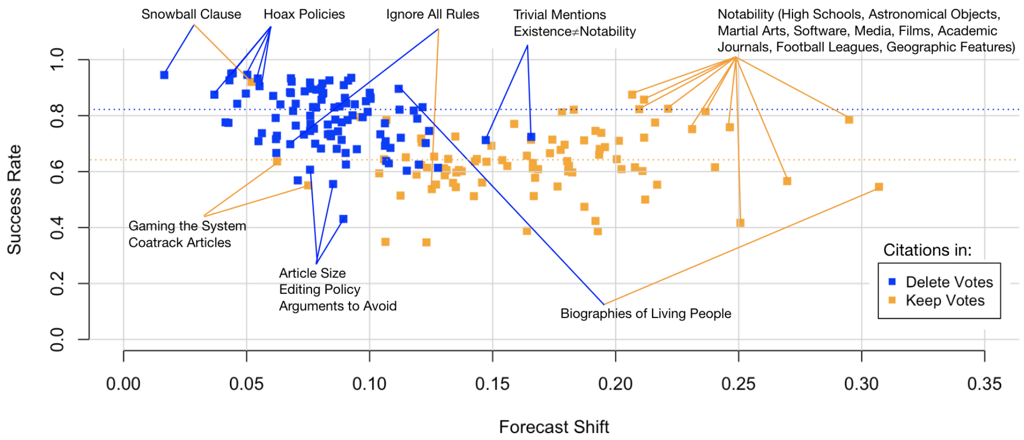


Fig. 12. Summary of success rates and forecast shifts for various policies. Scatter plot shows all policy pages with at least 25 citations in either `Keep` or `Delete` votes. Dotted lines mark baseline success rates.

contexts outside of *AfD* where no votes may be explicitly recorded. In particular, group decision-making research can benefit more broadly in analyzing conversations beyond the Wikipedia domain. We know that disparities in influence matter for group effectiveness. But getting at these patterns quantitatively is complex - most social science research instead avoids the question of extracting structure directly from text, instead relying on direct observable variables and survey data [31], or simulation with explicit preferences encoded in modeled agents [6]. We hope that the application of these methods will enhance any behavioral science research that studies groups and teams, and we believe our work gives a pathway towards quantifying successful and influential contributions in discussions more broadly in the future.

6.1 Next Steps for Analyzing *Articles for Deletion*

A significant limitation of our analysis in this work is that our forecast model is retrospective and correlational rather than causal. There are several ways that our research could be extended to use the same metrics but test specific hypotheses.

First, we are well-positioned now to measure discussion *quality* for the first time since early work in [40]. In that study, researchers identified poor decisions as those that were reversed at a later date: they flagged poor decisions either when an article was successfully re-nominated for deletion a previously kept article, or when a page that had previously been deleted as part of the *AfD* process was recreated. The magician from Figure 11, for instance, now has a recreated Wikipedia page with additional content. With an objective measure of whether articles “deserved" to be deleted, we can begin to eliminate the possibility of votes that were correlated merely with high- or low-quality articles rather than influence or impact on discussions. Another experiment would be to limit our analysis to close decisions with narrow margins, such as the 7.6% of votes ending in ties, or 5.2% of cases where administrators overruled the majority vote of participants. By limiting our scope to those discussions, we may be able to more narrowly test rhetorical strategies alone, rather than conflate rhetoric with a group’s overall sense of article quality.

But it is not obvious which action is appropriate to take when administrator decisions disagree with predictions from forecasts, opening a broader question of trust in machine learning systems. Can this model be used to recognize when a poor decision is being made, or when participating editors are missing key experience levels or subject matter expertise? Lam et al. [40] have already shown that diverse groups of decision-makers improves quality. Future implementations of forecast models in practice for Wikipedia could recommend either a pause in decision-making when a “surprising" outcome is being chosen by an administrator, or could even be extended to active recruiting of new voices that are potentially under-represented in existing discussion. New users might also be supported by direct recommendation of effective, narrowly scoped policies to consult when making a contribution, as identified by their forecast shift and success rates. This type of direct intervention will be a fruitful avenue for future work, though will need to be tempered by the unsteady reception to bots in the Wikipedia editorial system in general [19].

By going beyond raw statistics and into more granular, informed measurements of productivity, Wikipedia has an opportunity to greatly improve the measurement of quality, influential participation in their community. Moreover, the use of machine learning tools powered by natural language processing has precedent in that community: tools *already* exist and are in widespread use for numerous behind-the-scenes tasks like vandalism detection [53], bot detection [25], and article quality assessment [22]. We show that certain policies, especially `Notability` subpolicies, can be associated both with forecast shift and success and are used by effective inclusionist voters. The next step for this research is to recognize and describe the role of policy citation and related discourse behaviors in gatekeeping, enforcing or even intensifying advantages for long-term users. Our methodological framework is easily extended to evaluations based on aspects of a user’s profile, such as their self-identified race, gender, or interests. We have only scratched the surface of these topics, and even then have only begun to analyze *AfD*.

For tool developers interested in extending those findings to interventions, we believe our work is most promising as a catalyst for identifying the right set of voices for a discussion, pointing users at relevant discussions early to improve decision quality. In this work we have given detail on how early votes set the stage for later discussion in *AfD*, and how these shifts are larger when the first vote is contrary to the most common outcome. Setting aside the question of causality in this particular case, we know that priming effects are able to shape risk profiles, preferences, and topics under scrutiny in decision-making tasks [12, 32]. In that context it makes sense that decisions are driven by early participants, and that automated recommendation tools can be a plausible part of an improved community in the future.

In the *AfD* context, with the advantage of full discussion logs and explicit votes and outcomes, it also may be preferable to introduce other metrics that take more advantage of discussion structure, like the hypergraph representation from Hua et al. [28]. Taking advantage of this deeper structure, we can also use our forecast model for deeper temporal analysis of Wikipedia’s evolution over time, including a test of which policies have risen and fallen in prominence, success rates, and forecast shifts throughout the site’s rise and decline. This retrospective analysis is less focused on intervention in live settings than tools would be, but may be just as important for the community to understand itself and its own past.

REFERENCES

[1] David Adamson, Gregory Dyke, Hyeju Jang, and Carolyn Penstein Rosé. 2014. Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education* 24, 1 (2014), 92–124.

[2] Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling Deliberative Argumentation Strategies on Wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).* 2545–2555.

[3] Pablo Aragón, Vicenç Gómez, David García, and Andreas Kaltenbrunner. 2017. Generative models of online discussion threads: state of the art and research challenges. *Journal of Internet Services and Applications* 8, 1 (2017).

[4] Moira Burke and Robert Kraut. 2008. Mopping up: Modeling Wikipedia Promotion Decisions. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW).* ACM, 27–36.

[5] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1101–1110.

[6] Francisco Chiclana, JM Tapia Garciá, Maria Jose del Moral, and Enrique Herrera-Viedma. 2013. A statistical comparative study of different similarity measures of consensus in group decision making. *Information Sciences* 221 (2013), 110–123.

[7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[8] Simon DeDeo. 2014. Group minds and the case of Wikipedia. *Human Computation* (2014).

[9] Katie Derthick, Patrick Tsao, Travis Kriplean, Alan Borning, Mark Zachry, and David W McDonald. 2011. Collaborative sensemaking during admin permission granting in Wikipedia. In *International Conference on Online Communities and Social Computing*. Springer, 100–109.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL HLT Conference*.

[11] William Emigh and Susan C Herring. 2005. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the Hawaii International Conference on System Sciences*. IEEE.

[12] Hans-Peter Erb, Antoine Bioy, and Denis J Hilton. 2002. Choice preferences without inferences: Subconscious priming of risk attitudes. *Journal of Behavioral Decision Making* 15, 3 (2002), 251–262.

[13] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9, Aug (2008), 1871–1874.

[14] Heather Ford and Judy Wajcman. 2017. 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social Studies of Science* 47, 4 (2017), 511–527.

[15] Andrea Forte and Amy Bruckman. 2005. Why do people write for Wikipedia? Incentives to contribute to open–content publishing. *Proceedings of the ACM International Conference on Supporting Group Work* 5 (2005), 6–9.

[16] Andrea Forte and Amy Bruckman. 2008. Scaling consensus: Increasing decentralization in Wikipedia governance. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*. IEEE, 157–157.

[17] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72.

[18] Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet* 7, 3 (2015), 319–339.

[19] Richard Stuart Geiger. 2015. *Robots. txt: An Ethnographic Investigation of Automated Software Agents in User-Generated Content Platforms*. Ph.D. Dissertation. University of California, Berkeley.

[20] Richard Stuart Geiger and Heather Ford. 2011. Participation in Wikipedia's article deletion processes. In *Proceedings of the International Symposium on Wikis and Open Collaboration*. ACM, 201–202.

[21] Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature* 438 (2005), 900–901. Issue 7070.

[22] Aaron Halfaker. 2017. Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect. In *Proceedings of the International Symposium on Open Collaboration*. ACM, 19.

[23] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013).

[24] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the International Symposium on Wikis and Open Collaboration*. ACM, 163–172.

[25] Andrew Hall, Loren Terveen, and Aaron Halfaker. 2018. Bot Detection in Wikidata Using Behavioral and Other Informal Cues. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 64.

[26] Eszter Hargittai and Aaron Shaw. 2015. Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, Communication & Society* 18, 4 (2015), 424–442.

[27] Bradi Heaberlin and Simon DeDeo. 2016. The evolution of Wikipedia's norm network. *Future Internet* 8, 2 (2016), 14.

[28] Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2818–2823. http://aclweb.org/anthology/D18-1305

[29] Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. 2018. Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 74.

[30] Ali Javanmardi and Lu Xiao. 2019. What's in the Content of Wikipedia's Article for Deletion Discussions?. In *Proceedings of The Web Conference (WWW Companion)*. 1215–1223.

[31] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. 1999. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly* 44, 4 (1999), 741–763.

[32] Eric J Johnson, Suzanne B Shu, Benedict GC Dellaert, Craig Fox, Daniel G Goldstein, Gerald Häubl, Richard P Larrick, John W Payne, Ellen Peters, David Schkade, et al. 2012. Beyond nudges: Tools of a choice architecture. *Marketing*

Letters 23, 2 (2012), 487–504.

[33] Elisabeth Joyce, Jacqueline C Pike, and Brian S Butler. 2013. Rules and roles vs. consensus: Self-governed deliberative mass collaboration bureaucracies. *American Behavioral Scientist* 57, 5 (2013), 576–594.

[34] Dan Jurafsky and James H Martin. 2014. *Speech and language processing*. Vol. 3. Pearson London.

[35] Brian Keegan and Casey Fiesler. 2017. The Evolution and Consequences of Peer Producing Wikipedia's Rules. *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)* (2017).

[36] Brian Keegan and Darren Gergle. 2010. Egalitarians at the gate: One-sided gatekeeping practices in social media. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)*. 131–134.

[37] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 88.

[38] 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 (CHI)*. ACM, 453–462.

[39] Shyong K Lam, Jawed Karim, and John Riedl. 2010. The effects of group composition on decision quality in a social production community. In *Proceedings of the ACM International Conference on Supporting Group Work*. ACM, 55–64.

[40] Shyong K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. WP: Clubhouse?: an exploration of Wikipedia's gender imbalance. In *Proceedings of the International Symposium on Wikis and Open Collaboration*. ACM.

[41] Andrew Lih. 2004. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism, 2004*.

[42] Elijah Mayfield and Alan W Black. 2019. Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making. In *Workshop on Natural Language Processing + Computational Social Science at the North American Association for Computational Linguistics*.

[43] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.

[44] Amanda Menking and Ingrid Erickson. 2015. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, 207–210.

[45] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 219–245.

[46] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. ACM, 839–848.

[47] Jin Mu, Karsten Stegmann, Elijah Mayfield, Carolyn Rosé, and Frank Fischer. 2012. The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning* 7, 2 (2012), 285–305.

[48] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. 2018. Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 137.

[49] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.

[50] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[51] Ignacio J Pérez, Francisco Javier Cabrerizo, Sergio Alonso, YC Dong, Francisco Chiclana, and Enrique Herrera-Viedma. 2018. On dynamic consensus processes in group decision making problems. *Information Sciences* 459 (2018), 20–35.

[52] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning* 3, 3 (2008), 237–271.

[53] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. 2017. Building automated vandalism detection tools for Wikidata. In *Proceedings of the International Conference on World Wide Web (WWW) Companion*. 1647–1654.

[54] Jodi Schneider, Bluma S Gelley, and Aaron Halfaker. 2014. Accept, decline, postpone: How newcomer productivity is reduced in English Wikipedia by pre-publication review. In *Proceedings of the International Symposium on Open Collaboration*. ACM, 26.

[55] Jodi Schneider, Alexandre Passant, and Stefan Decker. 2012. Deletion discussions in Wikipedia: Decision factors and outcomes. In *Proceedings of the International Symposium on Wikis and Open Collaboration*. ACM, 17.

[56] Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the Conference on Computer Supported Cooperative Work*. ACM, 1069–1080.

[57] Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology* 48, 6 (1985), 1467.

[58] Jennifer Stromer-Galley and Peter Muhlberger. 2009. Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy. *Political Communication* 26, 2 (2009), 173–192.

[59] Bongwon Suh, Gregorio Convertino, Ed H Chi, and Peter Pirolli. 2009. The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM, 8.

[60] Dario Taraborelli and Giovanni Luca Ciampaglia. 2010. Beyond notability. Collective deliberation on content inclusion in Wikipedia. In *IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop (SASOW)*. 122–125.

[61] Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*.

[62] Wendy P Van Ginkel and Daan van Knippenberg. 2008. Group information elaboration and group decision making: The role of shared task representations. *Organizational Behavior and Human Decision processes* 105, 1 (2008), 82–97.

[63] Fernanda B Viegas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. 2007. Talk before you type: Coordination in Wikipedia. In *Proceedings of the Hawaii International Conference on System Sciences*. IEEE, 78–78.

[64] Lu Xiao. 2018. A Message's Persuasive Features in Wikipedia's Article for Deletion Discussions. In *Proceedings of the 9th International Conference on Social Media and Society*. ACM, 345–349.

[65] Lu Xiao and Nicole Askin. 2014. What influences online deliberation? A Wikipedia study. *Journal of the Association for Information Science and Technology* 65, 5 (2014), 898–910.

[66] Lu Xiao and Jeffrey Nickerson. 2019. Imperatives in Past Online Discussions: Another Helpful Source for Community Newcomers?. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[67] Lu Xiao and Niraj Sitaula. 2018. Sentiments in Wikipedia Articles for Deletion Discussions. In *International Conference on Information*. Springer, 81–86.

[68] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2000–2010.

[69] Diyi Yang, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia.. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*. 446–455.

A APPENDIX

A.1 Corpus Preprocessing

We evaluate an offline database of all Articles for Deletion discussions from the January 1, 2019 snapshot of Wikipedia, located at `https://dumps.wikimedia.org`. Compared to the broader internet, Wikipedia is simpler to preprocess due to the rigid formality of the archival process, the MediaWiki markup language, and enforced community standards. For most tasks, we are able to extract names, timestamps, and labels with only regular expressions.

A.1.1 Extracting Timestamps. AfD discussion norms require that all contributions are signed using a standard format, which includes the contributor's username or IP address and a timestamp in UTC format[14]. All lines following the outcome are checked for timestamps in Wikipedia standard format[15]:

```
\d\d:\d\d, \w+ \d+, 20\d\d (UTC)
```

A.1.2 Extracting outcomes. AfD discussions are archived in a specific format with only minor variation, and can be easily extracted for structured representation. We define a discussion as having an *outcome* if its archival page includes a header line with one of three fixed phrases (ignoring whitespace):

```
The result of the debate was [x]
The result was [x]
The result of this discussion was [x]
```

We save the captured string **`[x]`** as the debate outcome. When these lines are timestamped, we also log the user and timestamp of the outcome.

A.1.3 Extracting nominations, votes, and comments. If a timestamped contribution appears at the top of the discussion, prior to any votes, it is treated as a *nomination*. These statements have become more common over time: while they occur in only 67% of nominations in 2005, they were rapidly adopted and are present in 98% of nominations since 2008[16].

Following the nominating statement, any timestamped line is captured as either a vote or a comment. We define votes as any timestamped line beginning with a bolded phrase, following Wikipedia convention for contributions:

```
* ‘‘‘[y]’’’
```

Posts beginning with one or more leading asterisks creates a bulleted, threaded discussion. Words or phrases surrounded with three apostrophes creates ′′′**bolded**′′′ text. The value of this bolded text `[y]` is captured and stored. If no bolded phrase is present, but the line is still signed and timestamped, that line is treated as a *comment*[17]. Lines with no timestamped signature are discarded.

Several alternative solutions to deletion exist; each maintains the content of the page while deleting the page itself. In the five-label case, `Merge` and `Redirect`, the two most common alternate outcomes, are represented separately in line with prior work; in the two-label case they are merged

[14] These signatures are highly formulaic and easy to extract, because they can be automatically generated by MediaWiki's ~~~~ shorthand. When users do not sign contributions, bots add them, along with a citation to the `SIGNATURES` policy.

[15] In regular expressions, `\w` matches any letter and `\d` can match any numeric character. A `+` suffix captures one or more consecutive characters of that type.

[16] Under present policy, omitting a nominating statement is an acceptable reason for "speedy" dismissal and default "Keep" outcome for an AfD nomination.

[17] Lines beginning with the bolded phrase **"Comment"** are also treated as comments. Lines beginning with **"Note"** are automatically generated, typically for categorizing discussions by topic, and are discarded. Lines with "Relist" bolded are administrative notes to keep the discussion open for longer than the typical seven days, and are also discarded.

in with Delete. All other values are grouped together as `Other` in the five-label case[18]; in the two-label case they are merged in with Keep. Votes and outcomes of “Close", “Withdraw", and “Cancel" are treated as “Keep" outcomes as the page as well as its content is fully maintained. Copyright violations are treated as a “Delete" outcome, as the content is deleted as a result of the outcome. Any given vote or outcome is represented as a set that can contain zero or more normalized labels. Therefore, the probability of a vote for a particular label is not drawn from a distribution; probabilities of each label in L are disjoint.

A.1.4 Extracting users. For each nomination, outcome, vote, or comment, we log the user whose signature immediately appears before the timestamp, either with a MediaWiki link to their User page or their User Talk page:

```
[[User Talk:[z]
[[User:[z]
```

We extract **`[z]`** as a username and associate it with the nomination, outcome, vote, or comment where it was captured. When user signatures link to both User and User Talk pages and those usernames differ, the Talk page’s username is prioritized.

A.2 Corpus Format

We store all information about our corpus in JSON format.

Discussions are stored as a unique ID beginning with digit 1 and a title.

```
{
    "ID": 100300050,
    "Title": "Raymond Daniels"
}
```

Users are given a unique ID beginning with digit 2 and only their username is stored. Future work will include other demographic and metadata including time of registration and self-disclosed details from profile pages.

```
{
    "ID": 200002885,
    "Name": "Casliber"
}
```

Discussion outcomes are stored with a unique ID beginning with 3, a link to the parent discussion, the normalized set of outcomes from a small standardized set of options, the raw text as extracted from between the **“‘bold”’** quote marks, the unique ID of the user that posted the outcome, the Unix timestamp of the post, and the full text of the original outcome post. In general, signatures are stripped, though this is not perfect when users customize their signature formatting.

```
{
    "ID": 300272648,
    "Parent": 100300050,
    "Label": "keep",
    "Raw": "keep",
    "User": 200002885,
    "Timestamp": 1214101500,
   "Rationale": "The result was '''KEEP''' per [[WP:SNOW]], and meets guidelines.
                 Cheers, "
}
```

[18] “Userfy", “Transwiki", “Move", and “Incubate"

Votes match this formatting almost exactly, with unique IDs starting with 4 and links to the source discussion. An unused Parent field will link votes and comments hierarchically in a tree structure in future work. Non-voting comments follow the same structure, but have IDs beginning with 5 and do not contain the Label and Raw fields. Discussion-initial nominations also follow this structure, with IDs beginning with 6 and missing the same fields - their preferred outcome is presumed to be “Delete" as they are the nominating user.

```
{
    "Parent": -1,
    "Discussion": 100300050,
    "Timestamp": 1214094060,
    "User": 200009309,
    "Label": "keep",
    "Raw": "keep",
    "Rationale": "*'''Keep''' he is [[WP:V|verifiable]] and [[WP:N|notable]], as
                proven by the [[WP:RS|reliable sources]] cited in the article. ",
    "ID": 401391857
}
```

Each contribution has a corresponding entry in a dictionary labeled Citations. In these entries, each citation from user rationales is normalized from links and acronyms to the canonical name of the policy, guideline, or essay that is being cited.

```
{
    "ID": 401391857,
    "Citations": [
        "verifiability",
        "notability",
        "identifying reliable sources"
    ]
}
```

Received April 2019; revised June 2019; accepted August 2019