NORTHWESTERN UNIVERSITY

Socio-technical Systems for Identifying Latent Knowledge Gaps

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Computer Science and Communication Studies

By

Jim Maddock

EVANSTON, ILLINOIS

March 2022

ABSTRACT

Socio-technical Systems for Identifying Latent Knowledge Gaps

Jim Maddock

Asymmetric relationships between creators and consumers in peer-produced knowledge repositories produce inequitable knowledge representation–or knowledge gaps. These gaps result in unequal access to information, and downstream technologies that leverage peer-produced data perpetuate these inequities. Effective knowledge gap identification represents a necessary first step towards equitable knowledge representation. However, while prior work has uncovered a few important biases (e.g. gender, political, and cultural bias), no comprehensive and systematic way for identifying knowledge gaps exists.

In this dissertation we investigate current approaches for known knowledge gap mitigation, and we propose novel methods for latent knowledge gap identification through two studies. In other words, 1) how do editors currently address known unknowns, and 2) how do we identify unknown unknowns?

In our initial study we interview Wikipedia's editor community in order to better understand existing methods for knowledge gap identification. Study 1 documents editors' definitions of knowledge gaps, potential causes of knowledge gaps, and the social and

technical framework editors use to identify missing subjects and to create new content. We show that editors use a system of lightweight markers in order to distribute work throughout the community and to systematically "fill in" certain topical areas that are traditionally underrepresented. Ultimately, we argue that new technical systems need to leverage these existing social and technical frameworks–not rely on the creation of new workflows–in order to be successful. Our findings from Study 1 reinforce much of the existing empirical work on knowledge gaps, but represent a unique perspective grounded in the editor community.

Study 2 investigates one potential method for latent knowledge gap identification. In Study 2 we examine a *reader-sourced* approach, which leverages knowledge from Wikipedia's reader community in order to identify new knowledge gaps. We build on data produced by Wikipedia's Article Feedback Tool (AFT). Study 2 finds that, while it is challenging to build a machine classifier that can perfectly predict whether reader feedback will be helpful or unhelpful, we can still reduce editor workload associated with triaging reader feedback.

Acknowledgements

As with any dissertation, the complete list of individuals who deserve acknowledgement are too many to enumerate. To my sister, Jenna; my parents; my friends; my advisor, Darren, and my committee, Aaron, Aaron, and Ed; my many mentors in academia and industry along the way, my program colleagues and labmates, and the many Wikipedians who contributed to this work: thank you.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Wikipedia is one of the world's largest information reference works and publicly accessible knowledge repositories. As of December 2021, the English language version alone contained over 6.4 million articles.[1] In total Wikipedia spans 325 different language versions,[2] which include 57.7 million content pages. It is perhaps indicative of Wikipedia's ubiquity and preeminence as a reference source that the self-referential English Wikipedia article about Wikipedia[3] is likely the world's most reliable source on the subject.

Over the past year English Wikipedia received 125 billion page views, and Wikipedia.org was the thirteenth most visited domain globally by Alexa ranking.[4] Google search results include data collected from Wikipedia in both its answer box and knowledge panel, and increasingly common services such as Amazon's Alexa, Apple's Siri, and Google's Assistant query Wikipedia to answer users' questions, none of which is captured by the already massive pageview figure [**McMahon et al., 2017**]. Furthermore, downstream technologies such as Google's Knowledge Graph and a host of other AI systems leverage Wikipedia data; as Hovy et al. state in their 2013 overview, "not only are semi-structured resources [such as Wikipedia] enabling a renaissance of knowledge-rich AI techniques,

[1]Unless otherwise noted, all Wikipedia stats were collected from `https://stats.wikimedia.org/` on December 1, 2021.
[2]Language versions are also called "language editions" or just "Wikipedias"
[3]`https://en.wikipedia.org/wiki/Wikipedia`
[4]`https://www.alexa.com/topsites`

but...significant advances in high-end applications that require deep understanding capabilities can be achieved by synergistically exploiting large amounts of machine-readable structured knowledge [which is also provided by Wikipedia]" [**Hovy et al., 2013**].

Given the sheer number of humans and machines that consume information from Wikipedia, it is critical that knowledge is represented equitably. If Wikipedia's content is biased towards specific points of view–either through omission of certain topics or systematically skewed content–downstream consumers will likely perpetuate those biases. As the Wikimedia Foundation notes as justification for accurate and equitable knowledge representation, "readers often question the reliability of the content we create, notably because it is not accurate, not comprehensive, not neutral, or because they don't understand how it is produced, and by whom."[5] More insidiously, if readers do not question the reliability of systematically biased content, those biases become entrenched as fact.

1.1. Diverse Readers, Homogeneous Contributors

Wikipedia is a peer produced resource. While the Wikimedia Foundation–a non profit organization of around 500 employees–develops and hosts the technological infrastructure and platform that supports Wikipedia, the encyclopedia's content is wholly produced by unpaid volunteers, also called editors. Anyone with an internet connection and the technical skill to write Wikitext[6] can contribute. As of December 2021, English Wikipedia had been built by roughly 441,000 thousand editors, 41,000 of whom were considered active.

[5] `https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction#Reasoning:_Why_we_will_move_in_this_strategic_direction`

[6] Wikitext is a markup language used to create Wikipedia pages. More information can be found at `https://en.wikipedia.org/wiki/Help:Wikitext`.

The implications of Wikipedia's production model are complex. The sheer number of contributors and the quantity contributor hours invested in content production and maintenance have allowed Wikipedia to grow far larger than traditional print based encyclopedias while maintaining a similar or higher degree of reliability. For instance, in a highly cited but aging Nature article from 2005–now nearly two decades old–Jim Giles found that Wikipedia and Encyclopedia Britannica Articles were of similar quality [**Giles, 2005**]. But, while Wikipedia's breadth and quality can be attributed to it's volunteer contributor base, its decentralized governance structure means that content generally reflects the interests of its contributors. As long as their work follows Wikipedia's content guidelines, editors choose what content to produce. As we show in Chapter 3, editors tend to work on what they already know.

As web usership becomes more diverse, existing peer produced knowledge repositories (such as Wikipedia) under-represent their growing user base. Content creators remain heavily Western, Caucasian, and male, even though the fastest growing demographics of internet users come from the Global South. The organization Whose Knowledge reports that "[as of 2018, we know that 75% of the world's online population is from the global South, and 45% of all women in the world are online",[7] yet academic work continues to find that contributor demographics skew heavily towards the traditionally powerful [**Redi et al., 2021**]. This asymmetric relationship between creators and consumers can lead to inequitable knowledge representation [**Hecht and Gergle, 2009**]–or knowledge gaps–which is perpetuated in downstream technologies (e.g. semantic relatedness) that leverage peer produced data [**Hecht and Gergle, 2010a, Hovy et al., 2013**].

[7] `https://whoseknowledge.org/initiatives/decolonizing-the-internet/`

Prior work indicates a relatively prevalent self-focus bias among Wikipedia content, wherein editors tend to produce content that is relevant to their cultural or geographic context [**Hecht and Gergle, 2009**]. Therefore, in order to provide equitable knowledge representation, these systems need to actively support "communities that have been left out by structures of power and privilege" [**Zia et al., 2019c**]. Indeed, the Wikimedia Foundation's strategic direction for 2030 identifies knowledge gaps as a primary focus of research and development in the next five years[**Zia et al., 2019c**]. In other words, if Wikipedia aims to be "a comprehensive written compendium that contains information on all branches of knowledge"[8] it cannot only contain information of interest to its comparatively small and homogeneous population of editors.

1.2. Uncovering Latent Knowledge Gaps

While prior work has uncovered a few important knowledge gaps, no comprehensive and systematic way for identifying *latent knowledge gaps* exists. We define a latent knowledge gap as:

> ***Latent Knowledge Gap:*** *A content gap that is undocumented by prior content gap research and invisible to the editor community. Broadly a* known *knowledge gap is a "known unknown", whereas a* latent *knowledge gap is an "unknown unknown".*

We note that the phrase *to the editor community* represents a key component of our definition. Identifying and filling knowledge gaps that are unknown to all of humanity falls under the purview of general research, not bias mitigation. While Wikipedia aims to

[8] `https://en.wikipedia.org/wiki/Wikipedia:Purpose`

contain the sum total of human knowledge, the encyclopedia's policies explicitly exclude original, unpublished research. For example, under this definition we would not consider a new discovery in physics or chemistry to represent a latent knowledge gap until after its publication in a peer reviewed source. In other words, in order to merit inclusion a subject must be "known"–e.g. researched and verified–to some subset of the human population. We therefore define *latent knowledge gaps* as an "unknown unknown" only to those with the capacity to add information to Wikipedia. We use this scoped definition throughout the document.

As one example of a known knowledge gap, prior work shows female historical figures are represented differently than male historical figures on Wikipedia. In this case researchers targeted the knowledge gap *a priori* and attempted to characterize its scope (e.g. [**Wagner et al., 2015**]). Since the start of this dissertation, the Wikimedia Foundation has published a "Taxonomy of Knowledge Gaps" [**Redi et al., 2021**] which amounts to a meta-study of knowledge gap research. However, while the taxonomy provides a thorough overview of known gaps, it does not provide methods for uncovering new gaps.

Investigation and quantification of known gaps represents an important component of attaining more equitable knowledge representation. Quantification illustrates *how asymmetrically* Wikipedia covers specific topical areas, and it allows editors or organizations to direct effort and attention towards these topics. For instance, ongoing analysis of Wikipedia's gender gap provides an empirical foundation for the project Women in Red,[9]

[9]Women in Red's (`https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red`) mission statement reads: "We are a group of editors of all genders living around the world focused on reducing systemic bias in the wiki movement. We recognized a need for this work as, in October 2014, only 15.53% of English Wikipedia's biographies were about women.[**Graells-Garrido et al., 2015**]" (citation reproduced from the original text)

which has been remarkably successful at adding missing women's biographies, countering systematic bias against women, and advocating for policy changes that continue to marginalize underrepresented groups. However, exclusive focus on *a priori* approaches to addressing content asymmetries becomes problematic because editors cannot fix *latent* knowledge gaps, or gaps that they do not know to look for. While high profile, known biases–e.g. gender bias [**Wagner et al., 2015**], political bias [**Greenstein and Zhu, 2012**], and cultural bias [**Hecht and Gergle, 2010a**]–have been studied extensively, a profusion of other biases undoubtedly exist that remain invisible to research and editor communities.

In this dissertation we investigate current approaches for known knowledge gap mitigation, and we propose novel methods for latent knowledge gap identification through two studies. In other words, 1) how do editors currently address known unknowns, and 2) how do we identify missing information that editors may not know to look for? In our initial study (the right-most section of Figure 1.1), we interview Wikipedia's editor community in order to better understand existing methods for knowledge gap identification. Using an asynchronous remote community of 19 editors, we investigate the following research questions:

> ***RQ 1.a:*** *What processes do editors currently use to identify and mitigate knowledge gaps?*
>
> ***RQ 1.b:*** *How can socio-technical systems better facilitate existing knowledge gap identification and mitigation processes?*

Study 1 documents editor's definitions of knowledge gaps, potential causes of knowledge gaps, and the social and technical framework editors use to identify missing subjects

and to create new content. We show that editors use a system of lightweight markers in order to distribute work throughout the community and to systematically "fill in" certain topical areas that are traditionally underrepresented. Ultimately, we argue that new technical systems need to leverage these existing social and technical frameworks–not rely on the creation of new workflows–in order to be successful. Our findings from Study 1 reinforce much of the existing empirical work on knowledge gaps, but represent a unique perspective grounded in the editor community.

Study 2 (the central component of Figure 1.1) investigates one potential method for latent knowledge gap identification. In study 2 we examine a *reader-sourced* approach, which leverages knowledge from Wikipedia's reader community in order to identify new knowledge gaps–i.e. content gaps that are not obvious to the population responsible for producing Wikipedia's content. Theoretically, this approach would allow editors to gather information about missing content from Wikipedia's readers–a community which is much broader and more diverse than Wikipedia's editor community–which could in turn diversify content production. We build on the reader-sourced approach and data produced by Wikipedia's Article Feedback Tool (AFT) [**Halfaker et al., 2013**], [10] and we discuss the theoretical underpinnings for this method in Chapter 2.

A necessary step in implementing a reader-sourced system is sorting helpful and unhelpful feedback. The initial version of the AFT gathered a large quantity of unhelpful feedback from readers, which in turn created extra work for editors as they triaged the resulting backlog of comments. As we show in Study 1, editors have an already limited amount of time to allocate towards the functionally boundless task of improving

[10]For an overview of the Article Feedback Tool, see: `https://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool`.

Wikipedia, and the extra work required to use the AFT lead to the system's ultimate abandonment.[11] Study 2 aims to address this fundamental flaw by addressing the following research questions:

> ***RQ 2.a:*** *Can we classify, sort, rank, and present reader feedback to make it most useful to editors?*
>
> ***RQ 2.b:*** *To what extent can we reduce overall workload associated with processing reader feedback?*

Study 2 finds that, while it is challenging to build a machine classifier that can perfectly predict whether reader feedback will be helpful or unhelpful, we can still reduce editor workload associated with triaging reader feedback. By only making high confidence predictions we can maintain relatively high precision for both helpful and unhelpful predictions while reducing the overall quantity of triage work required of editors. Ultimately, we find that our classifier filters unhelpful feedback in the form of spam and vandalism more effectively than it uncovers uniquely helpful feedback, likely due to the highly context dependent nature of what makes feedback helpful.

In many ways, this dissertation aims to provide both *practical* and *feasible* solutions to addressing Wikipedia's knowledge gaps at the expense of eschewing broad theoretical claims or generalizability to other platforms. We motivate both studies with existing theoretical and empirical work where applicable–and in the case of Study 1 we highlight areas where our findings bring together multiple areas of prior research. But our ultimate goal is to provide solutions to knowledge gap identification that are grounded in data

[11]For archives of editor commentary about the AFT, see: `https://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/Article_feedback` and `https://en.wikipedia.org/wiki/Wikipedia_talk:Article_Feedback_Tool/Version_5/Archive_1`. We discuss the AFT more extensively in Chapter 2

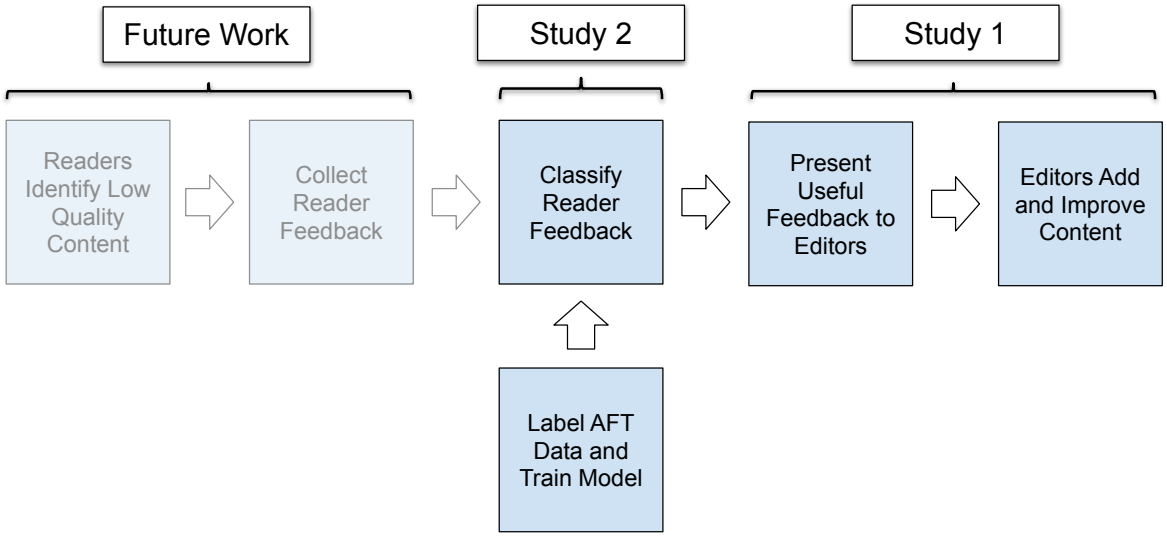

Figure 1.1. A visual representation of how Studies 1 and 2 fit together. Study 2 provides a method for collecting data from readers, while Study 1 provides an understanding of how to integrate that data into existing workflows. Study 2 uses data collected by the AFT, but a real world system would collect and classify new feedback from readers.

from those who are doing the day to day work and represent our metaphorical "boots on the ground": Wikipedia's editor community. This Wikipedia-specific approach may limit generalizability to other domains, but it does so in order to provide concrete, actionable solutions for humanity's single largest knowledge repository. Future studies could potentially extend these findings to other platforms, but that work is beyond the scope of this dissertation.

We organize the remainder of this dissertation as follows: Chapter 2 provides an overview of prior and existing knowledge gap identification strategies, as well as reader-sourcing's theoretical foundation in peripheral participation. Chapter 3 describes "Characterizing Existing Practices for Identifying and Mitigating Knowledge Gaps" (Study 1),

which used an asynchronous remote community to explore editors' existing methods for identifying knowledge gaps and creating new content, aiming to understand how new socio-technical systems could integrate and augment these workflows. Chapter 4 describes "Classifying Reader-Sourced Feedback for Knowledge Gap Identification" (Study 2), which uses a machine classifier to improve one potential method of knowledge gap identification. Chapter 5 synthesises our findings from Studies 1 and 2 and offers several possible directions for future work.

CHAPTER 2

Background and Related Work

Knowledge gaps represent a critical issue to any comprehensive knowledge repository [**Zia et al., 2019c, Zia et al., 2019b, Zia et al., 2019a**], and may be especially prevalent in peer produced resources. Exact definitions of "knowledge gap" vary as we discuss extensively in Chapter 3, but the Wikimedia Foundation's research team provides the following definition:

> *[Knowledge gaps are] disparities in content coverage or participation of a specific group of readers or contributors. A gap corresponds to an individual aspect of the Wikimedia ecosystem—for example readers' gender, or images in content—for which we found evidence of a lack of diversity, or imbalanced coverage across its inner categories (for example, proportion of readers who identify as men, women or non-binary in the case of the reader gender gap).* [**Redi et al., 2021**]

This definition leads to some confusion, as it includes both *contributors* and *readers* under the umbrella of "knowledge gaps". In our experience (again, see Chapter 3) the definition of "Content gap" maps more closely to most individuals' conceptions of a knowledge gap:

> *In the widest sense, "content" is information about a topic, i.e. a piece of knowledge that could be the focus of one or more Wikipedia articles.*

> *"Coverage" refers to how well Wikimedia project content addresses a particular topic. In turn, a content gap refers to differences in coverage of one or more topics.* [**Redi et al., 2021**]

For the remainder of this document we use the terms "content gap" and "knowledge gap" interchangeably. When referring to asymmetries in Wikipedia's reader or contributor populations, we specify "reader gap" or "contributor gap", respectively.

2.1. Causes of Knowledge Gaps

Prior theoretical and empirical work indicates that peer produced resources exhibit a strong self-focus bias; contributors tend to focus on producing content or resources that are most relevant to their interests [**Hecht and Gergle, 2009, Hecht and Gergle, 2010a**]. For instance, in a study of 15 language editions, Hecht & Gergle find that articles related to a particular Wikipedia's "home country" have more inlinks than articles related to a different country. This finding makes intuitive sense. Due to the volunteer nature of peer-produced repositories, contributors choose to develop resources that are most relevant to their interests or background. More recent work by Das et al. complicates this narrative slightly, showing that Open Street Map contributors do not edit along strict gender lines [**Das et al., 2019**]. Male editors make more edits to traditionally feminized spaces, and vice versa. Nevertheless, this finding does not negate over a decade of Wikipedia specific research that investigates the relationship between self-focus bias and other contributor facets (e.g. cultural background).

Given a diverse contributor population, self-focus bias would not necessarily result in inequitable resource production and representation. However, in the case of Wikipedia–and many peer-produced knowledge repositories–the contributor population is relatively homogeneous [**Collier and Bear, 2012, Lam et al., 2011**]. Furthermore, the distribution of edits per editor follows a power law distribution, which concentrates most content production among an even smaller group of editors [**Ortega et al., 2008, Shaw and Hill, 2014**]. The combination of editor homogeneity and self-focus bias could theoretically result in systematic knowledge gaps–or downstream content bias–such as gender, political, and cultural biases.

Since proposing this dissertation, the Wikimedia Foundation has named this phenomenon the "Contributor Gap", and the research team has published an extensive meta-review of academic work, internal surveys, and community comments that aim to understand and quantify Wikipedia's lack of contributor diversity [**Redi et al., 2021**]. Broadly, they break the contributor gap into multiple dimensions (which they label "gaps"), including gender [**Hinnosaar, 2019, Protonotarios et al., 2016, Shaw and Hargittai, 2018**], age [**Hinnosaar, 2019, Protonotarios et al., 2016, Shaw and Hargittai, 2018**], geography language [**Johnson et al., 2016, Shaw and Hargittai, 2018**], socio-economic status [**Hinnosaar, 2019, Shaw and Hargittai, 2018**], sexual orientation [**Wexelbaum et al., 2015**], and cultural background [**Keegan, 2019, Shi et al., 2019, Shaw and Hargittai, 2018**]. Although reviewing the considerable quantity of prior work covered in their meta-review is outside the scope of this dissertation, each facet represents a dimension along which lack of contributor diversity could lead to inequitable content production.

Other causes of content gaps exist, though in many ways these causes are a byproduct of contributor homogeneity. For instance, policies that define Wikipedia's scope–i.e., which subjects are notable enough to merit inclusion–tend to bias the encyclopedia against articles about women and other traditionally underrepresented groups [**Wagner et al., 2016, Ford and Wajcman, 2017, Taraborelli and Ciampaglia, 2010**]. Furthermore, Wikipedia's sourcing guidelines specify that editors should prioritize peer reviewed, secondary sources, which effectively excludes large topical areas not extensively documented by western academics [**Ford et al., 2013, Gallert and Van der Velden, 2013**]. We explore these causes more extensively in Chapter 3.

2.2. Wikipedia and Traditional Encyclopedias

An expansive body of research has sought to understand whether Wikipedia does indeed suffer from knowledge gaps. Early in Wikipedia's lifespan, researchers compared both the coverage and quality of Wikipedia articles to analogous articles in traditional encyclopedias in order to understand the veracity of peer-produced knowledge [**Giles, 2005, Halavais and Lackaff, 2008, Holman Rector, 2008**]. These studies generally found Wikipedia's coverage to be comparable or better than traditional encyclopedias, especially among topics categorized as math or science, and Wikipedia's quality to be similar but less consistent.

Studies that compared Wikipedia to traditional encyclopedias used small sample sizes, ranging from tens of articles to thousands of articles. While this research may indicate that Wikipedia does not suffer from glaring knowledge gaps compared to traditional encyclopedias, the studies and articles examined are not comprehensive and therefore may not

be representative. Drawing conclusions about the completeness of Wikipedia's coverage from a small subset of articles would be inconclusive at best.

Furthermore, both Wikipedia and traditional encyclopedias may contain similar systematic biases. Though traditional encyclopedias are produced by domain experts, these experts likely encode their own self-focus bias. Early work indicates that Wikipedia contains more information on certain topics than traditional encyclopedias, which could be an indication of knowledge gaps in expert curated sources [**Halavais and Lackaff, 2008**]. Using comparative analysis would therefore miss latent biases that exist in both sources.

2.3. Quantifying Knowledge Gaps Identified A Priori

Researchers explore systematically underrepresented information (i.e. coverage bias) by comparing Wikipedia articles, topics areas, or language editions to each other. These projects define a type of bias a priori and attempt to quantify the extent of that bias. Prior work has compared Wikipedia's representation of male to female historical figures (gender bias) [**Wagner et al., 2015**], use of liberal to conservative terms in political articles (political bias) [**Greenstein and Zhu, 2012**], and representation of topics that are culturally relevant to a particular language edition to those that are not (cultural bias) [**Hecht and Gergle, 2010a**].

Results from these studies are mixed, indicating that systematic knowledge gaps do exist throughout Wikipedia, though not always in the form that researchers expect. For instance, Wagner and colleagues find no evidence of coverage or visibility bias with respect to gender, but they find strong evidence of structural and lexical bias [**Graells-Garrido**

et al., 2015, Wagner et al., 2015]. While articles about women are slightly over-represented compared to men, articles about women more frequently focus on romantic and family related issues and are more likely to be linked to men than vice versa. In contrast, Reagle & Rhue find that Wikipedia under-represents female historical figures when compared to male historical figures [**Reagle and Rhue, 2011**].

Studies that investigate political bias find that Wikipedia articles generally favor liberal viewpoints [**Greenstein and Zhu, 2016**], but that over time the encyclopedia has trended towards more equitable representation of liberal and conservative topics [**Greenstein and Zhu, 2012**]. Additionally, certain topics contain systematic political bias [**Greenstein and Zhu, 2012**]. For instance, articles about civil rights consistently favor language used by liberal politicians, and articles about trade favor conservative language.

Prior work that investigates cultural bias reveals that much, if not most, content is not universal across language editions [**Bellomi and Bonato, 2005, Hecht and Gergle, 2010a**]. For example, Hecht and colleagues show that the French Wikipedia article for "Conspiracy theory" contains information about the Algerian War and the English version does not. Services such as Omnipedia and WikiBrain that visualize these content asymmetries reveal that a majority of subject matter is only available in single language editions [**Bao et al., 2012, Sen et al., 2014**]. More recent work by He et al. shows that content asymmetries extend beyond written content, finding that Wikipedia's images are *more* diverse than text [**He et al., 2018**].

This body of work implies that other knowledge gaps likely exist. Given that Wikipedia contains all of these "searched for" biases in one form or another, there is no theoretical or empirical reason to believe that this list is comprehensive or exhaustive. Latent

knowledge gaps could exist, but these gaps will remain invisible without systematic identification strategies.

2.4. Approaches Towards Systematically Identifying Knowledge Gaps

One approach towards knowledge gap identification focuses on increasing editor diversity. If indeed self-focus bias is responsible for knowledge gaps, then diversifying the editor population would also diversify content production. A multitude of complex factors–including gatekeeping [**Bryant et al., 2005**], the internet skills gap [**Hargittai and Shaw, 2015**], and language barriers [**Hale, 2014**]–are likely responsible for editor homogeneity. Prior attempts to reduce these barriers through newcomer socialization programs [**Narayan et al., 2017**], edit-a-thons [**Lavin, 2016**], and third party organizations have been somewhat successful, however the population of editors has not grown substantially, and the majority of edits are still made by a comparatively small group of editors.

An alternate approach focuses on exposing knowledge gaps to the existing population of editors to effectively reduce self-focus bias. In the past researchers have successfully leveraged this approach to identify low quality articles. For instance, Warncke-Wang and colleagues built a machine learning model to predict an article's quality from structural features [**Warncke-Wang et al., 2015, Warncke-Wang et al., 2013**]. The Objective Revision Evaluation Service (ORES)[1] represents the production version of this work, which is maintained by Wikimedia for quality assessment multiple Wikis, including several Wikipedia language editions [**Halfaker and Geiger, 2020**]. Wikipedia editors use ORES to identify low quality areas of the encyclopedia, and they can allocate their attention to these areas.

[1] `https://www.mediawiki.org/wiki/ORES`

While automated systems such as ORES identify quality gaps without onboarding new volunteers, they suffer from two major drawbacks when applied to missing knowledge identification. First, automated systems require features to rank or grade articles. With quality assessment generating features turns out to be relatively straightforward; an article's structural features (e.g. citations, images, wikilinks, etc.) provide a fairly accurate proxy for quality. However, in the case of missing content, developing a feature-set becomes much more challenging. While a missing image or a low number of citations might indicate low article quality, it is not clear that similar structural features could identify missing information within an article, or how these features could be leveraged to identify an article that does not yet exist.

Structural features do not represent the entire range of features that could be used in an automated system, but other types of features can be expensive to extract or difficult to leverage effectively. For instance, early automated approaches used graph based or lexical features to assess Wikipedia's topical coverage [**Holloway et al., 2007**]. However, these systems saw limited development for practical applications, likely due in part to the disconnect between academic research and the Wikimedia Foundation's engineering priorities.

Second, automated systems can re-encode and reinforce bias when built with biased training data [**Halfaker and Geiger, 2020**]. This is less problematic when identifying low quality articles because Wikipedia has relatively objective criteria for judging article quality. However, if automated systems for knowledge gap identification are built using training data provided by editors, these algorithms could re-encode existing self-focus

bias. For instance, systems that leverage Wikipedia's category structure may not uncover latent gaps due to the way in which the categories themselves are defined.

2.4.1. Reader Sourced Systems

Reader-sourced systems represent a third approach to latent knowledge gap identification, and these systems provide the foundation for Study 2 (see Chapter 4). Reader-sourced systems blur the boundary between "reader" and "contributor", effectively reducing barriers to making a contribution. These systems allow readers to make small contributions while investing minimal effort to both create content and learn community norms. The Article Feedback Tool (AFT) represents one such effort to increase article quality through reader-sourcing. Developed by researchers at the Wikimedia Foundation, the tool was first launched in 2010 and discontinued after five iterations in 2014. The AFT presented readers with a simple suggestion form, as illustrated in Figure 2.1. Reader responses identified low quality articles and provided editors with suggestions for improving those articles [**Halfaker et al., 2013**].

Reader-sourcing has theoretical roots in legitimate peripheral participation [**Lave, 1991**]. Newcomers develop into contributors by first making small contributions or participating in peripheral aspects of the community. As these newcomers become more familiar with community norms, they take on more centralized roles. Preece and Shneiderman developed a framework for online communities based on legitimate peripheral participation, which argues that a community's users progress from readers to contributors and collaborators, and eventually they become leaders [**Preece and Shneiderman, 2009**]. At each step of the framework, fewer users progress to the next step; while a community may

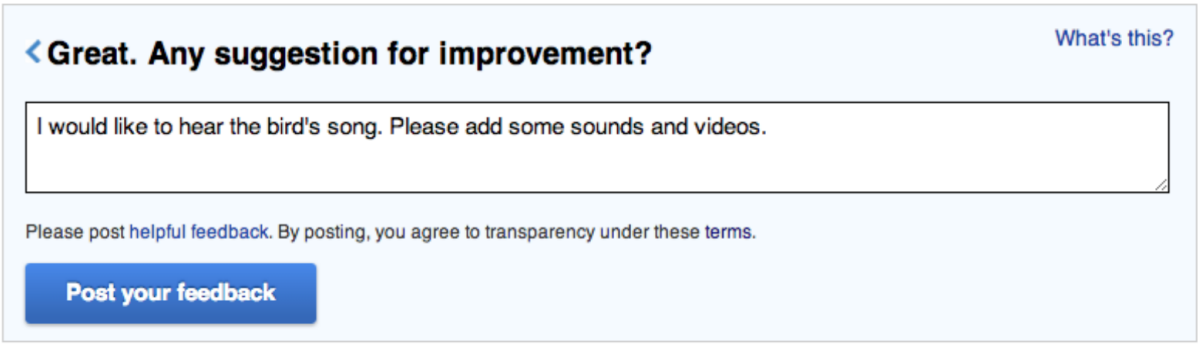

Figure 2.1. One version of the AFT's suggestion form, which allowed readers to provide feedback about Wikipedia articles. Reproduced from `https://upload.wikimedia.org/wikipedia/commons/d/d2/Article_Feedback_Slides_-_08-11-2013.pdf`

have many readers, few of those readers will develop into leaders. Scholars have leveraged Preece and Shneiderman's framework to argue that characterizing Wikipedia's readers as free-riders is inaccurate, and in many cases reading leads to contribution [**Antin and Cheshire, 2010**].

Reader-sourced systems offer one possible identification strategy for latent gaps due to the size and diversity of Wikipedia's readership. In an ideal scenario, a reader-sourced approach would effectively increase editor diversity and reduce the effect of self-focus bias, all while avoiding the costs and problems associated with recruiting and onboarding newcomers. With relatively low time investment, readers from diverse backgrounds could identify content systematically missing from Wikipedia. Experienced editors familiar with Wikipedia's community norms would then work to improve these areas of the encyclopedia, seeking appropriate expertise when needed. While editors may not possess the expertise to create or improve content related to certain knowledge gaps, acknowledging that the gap exists represents a necessary first step.

Developing an effective reader-sourced system faces a number of challenges. Existing editors have a finite amount of time to allocate towards maintaining and improving Wikipedia, and therefore will not adopt tools that increase workload. The AFT received push-back from the editor community due to its lack of integration with standard workflows and high volume of non-actionable and bad-faith reader feedback.[2] As a result, despite five different deployments the AFT was discontinued in 2014. Ultimately, any reader-sourced system for identifying knowledge gaps must address this point of failure.

The following chapters explore current approaches leveraged by editors for knowledge gap mitigation, and the viability of reader-sourced systems for knowledge gap identification. Study 1 in Chapter 3 looks generally at existing social and technical frameworks that support knowledge gap identification. While this study investigates editor conceptualizations of knowledge gaps more broadly, it also provides a foundational understanding of *where* and *how* reader sourced systems could integrate with existing workflows. Study 2 in Chapter 4 then specifically explores the possibility of using a machine classifier to filter and rank reader feedback in order to reduce the extra burden a reader sourced system would place on the editor community. Figure 1.1 provides a visual representation of how Studies 1 and 2 contribute to the overall goal of this dissertation.

[2]For archives of editor commentary about the AFT, see: `https://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/Article_feedback` and `https://en.wikipedia.org/wiki/Wikipedia_talk:Article_Feedback_Tool/Version_5/Archive_1`. The Wikimedia Foundation's decision to disband the AFT can be found here: `https://phabricator.wikimedia.org/T47538`

CHAPTER 3

Study 1: Characterizing Existing Practices for Identifying and Mitigating Knowledge Gaps

Over Wikipedia's multiple decade-long lifespan, researchers and community members have increasingly identified, quantified, and documented biases throughout the encyclopedia. One promising consequence of increased awareness has been the development of organizations and groups of editors that add and improve content in order to cultivate more equitable and balanced knowledge representation. As one example, the organization Whose Knowledge identifies as a "global campaign to center the knowledge of marginalized communities."[1] Over its five year lifespan, Whose Knowledge has led multiple initiatives to add missing images of influential women to Wikipedia, particularly focusing on influential women of color.

These communities represent an ideal starting point for understanding how editors conceptualize and approach knowledge gaps. Ultimately Wikipedia's editors–not algorithms or tools–produce the content that constitutes the encyclopedia. Though this may seem self-evident, editors often lament the disconnect between Wikipedia's producers and the engineers and researchers who develop interventions; interventions don't stick if they don't fit with the needs of the producers. As such, in order to support the communities

[1] `https://whoseknowledge.org/`

already engaged in mitigating Wikipedia's various biases, researchers must first engage with those communities in order to understand their workflows and processes.

The first phase of this dissertation used an Asynchronous Remote Community (ARC) study [**MacLeod et al., 2016**] of these editors to understand existing practices for identifying knowledge gaps. At a high level, ARC studies create a small, private, online community with participants from the population of interest, which allows researchers to collect data from geographically distributed individuals while facilitating focus-group-like interaction and between participants. Over the course of several weeks, participants participate in an online, asynchronous focus group. Using this methodology, study 1 attempts to answer three critical questions:

(1) **Identification of knowledge gaps:** *What process do editors use to identify missing information?* Existing practices can inform successful tool design. Editors may leverage processes for identifying knowledge gaps which could be augmented to work more effectively or scaled to work across the entire encyclopedia. Current research does not investigate the extent to which editors systematically identify gaps, if at all.

(2) **Triage:** *When editors identify a knowledge gap, how do they triage the missing information and determine which gaps most deserve attention?* Not all knowledge gaps are created equal, and editors have limited time and energy. Some missing content only leads to incomplete information, while systematically excluded content can lead to inequitable, under-representation. For instance, while lower numbers of articles about female historical figures would indicate problematic under-representation and gender bias, an incomplete article about Western

Military History (which is a traditionally over-produced topic [**Warncke-Wang et al., 2015**]) would not. Given limited time and energy, editors must choose to prioritize one knowledge gap over another. Critically, in this example over-production of Western Military History is related to the demographics of most Wikipedia editors, while women historical figures represent a minority interest. This connection between under-representation within the editor community and underproduction can lead to systematic knowledge gaps [**Redi et al., 2021**].

(3) **Editor Workflow Integration:** *How can a tool for knowledge gap identification augment and integrate into existing editor workflows?* Due to the amount of work required to maintain and improve Wikipedia, editors must allocate their attention efficiently. Editors must use their limited time to add new content, improve existing content, and police vandalism, in addition to performing a host of community building, management, and newcomer socialization tasks. Therefore, systems and tools that add to editors' workloads are not adopted or are quickly abandoned.[2] For example, while the AFT provided some useful information, lack of editor workflow integration and a high volume of non-actionable feedback required that editors spend large amounts of time searching for helpful suggestions. In order to avoid the adoption problems that befell prior reader-sourced systems, this project must integrate and support current practices for adding missing content.

This study aims to understand which processes exist for identifying and mitigating knowledge gaps. For instance, some editors might choose to allocate time to articles where

[2] `https://www.mediawiki.org/wiki/Article_feedback/Version_5/Report#Overview`

known gaps exist, some might add information to articles where they are subject experts, and others might actively look for new knowledge gaps to add to "Articles for Creation". Characterizing these different processes represents a necessary first step towards developing social and/or technical interventions that support editors in mitigating knowledge gaps.

3.1. Methods

We conducted a single four week study using ARC (described in detail below) with participants recruited from Wikipedia editor communities. We specifically targeted editors from communities that aim to identify and fill knowledge gaps, such as WikiProject Women writers/Missing articles[3] (a project that mitigates gender bias) and Whose Knowledge.[4]

3.2. Data Collection

We collected data for this study using an Asynchronous Remote Community (ARC). In prior work, researchers have used ARC to create focus groups with participants who are geographically distributed [**MacLeod et al., 2016, Maestre et al., 2018**]. As in a focus group, participants are able to interact with one another while responding to questions and prompts. However, because the study is conducted asynchronously within a remote community, participants do not need to be physically co-located, and they can respond on their own time.

[3] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_writers/Missing_articles`
[4] `https://whoseknowledge.org/`

For this research, the ARC provides one major benefit over traditional interviews. Interaction between participants creates potential for richer and deeper responses in situations where the researcher is not already embedded within the community of interest. While in a traditional interview the researcher might not possess enough expertise or experience to know which questions to ask, both focus groups and ARC encourage participants to respond to one another, building from each other's ideas. Additionally, ARC participants can refocus or reformulate prompts as well as provide feedback around topics, questions, or problems the researcher should consider that would otherwise remain unknown and undiscussed.

We hosted our online community using the platform Focus Group It,[5] which provides a Facebook-like user interface for posting and commenting.[6] The community was a private, invite-only group so that all participant data is not shared with others outside of the study. All 19 editors participated in the same online community and answered the same prompts.

In our screener we asked participants whether they would be comfortable talking about their Wikipedia edit history because–when possible–we asked participants to provide specific examples from their work. Editors' edit histories are already publicly available through change logs, but editors were required to be comfortable commenting on and explaining specific past experiences of their choosing. We include several of these specific examples in the results section of this study.

[5] `https://www.focusgroupit.com/`

[6] While the majority of prior work has used private Facebook groups, Northwestern's IRB no longer allows researchers to host communities through Facebook

We provided participants with an overview of the study methodology and required them to sign a code of conduct (see Appendix C)[7] before joining the online community. Generally, the code of conduct encouraged participants to be respectful of one another's opinions and viewpoints, to avoid antagonizing language, and to be considerate of different ethnic, socio-economic, and educational backgrounds. The code of conduct also outlined community moderation policies for cases that could not be resolved within the community, though no such cases arose during the study. Participants were permitted to withdraw from the study at any time, though none did so explicitly.

The study lasted 4 weeks. Participants were required to answer 2 prompts a week and comment on 2 responses for each prompt. All 8 prompts are included in Appendix E. The following is an example of a prompt we used in this study:

> *For your fifth post we would like you to think broadly about "knowledge gaps". How would you define a "knowledge gap" in the context of Wikipedia? You can reuse or modify the definition you gave in the screener survey. How are the experiences with missing content and/or bias you discussed in prior responses related to your definition? With this definition in mind, think about the specific challenges you face when addressing a knowledge gap. What is the biggest hurdle or obstacle you encounter? If you could implement something to aid this process (a tool, a policy change, a new collaborative effort) what would it be?*

Responses are participants' answers to a prompt. Most responses were written, but for the final 2 prompts participants were required to sketch a process flowchart

[7]Our code of conduct was based on [**Walker and DeVito, 2020**].

or diagram and upload this sketch to the community. Participants were instructed to "plan to write a few short paragraphs" for each prompt, but many participants wrote substantially more.

Comments are other participants' thoughts about a specific response.

Each prompt was designed to take 20 minutes, and each response to take 10 minutes, resulting in a minimum total time of about 60 minutes per week and about 4 hours of work over the course of 4 weeks. To facilitate conversation or follow up on specific points, researchers commented on specific responses throughout the study.

At the end of the study, participants were compensated with $100 for their time if they completed 75% of the prompts, and $125 if they completed 100% of the prompts. Participants were paid using Paypal or Amazon Gift Cards, depending on their country of residence and personal preference.

3.3. Overview of ARC Protocol

Study 1's prompts fell roughly into three parts, as outlined below. These parts represent the range of information we collected over the course of the study. All questions were answered and commented on within the ARC setting. For the final part (e.g. *workflows*), editors captured and uploaded the physical artifacts they created to the community.

(1) **Introduction and background:** The researcher explained the purpose of the project as a whole, and prompted participants to introduce themselves. The researcher asked for the participants' backgrounds and experience in editing Wikipedia, including why participants started editing, how participants describe their roles as editors, and which topics or pages the participants chooses to edit.

(2) **Knowledge gaps:** The researcher probed the editors' conceptions of knowledge gaps. Editors were asked how they would define a knowledge gap, and whether they could describe different types of knowledge gaps. The researcher also asked whether editors currently identify knowledge gaps, and if so to describe their process. Finally, editors described their criteria and processes for deciding which knowledge gaps to fill.

(3) **Workflows:** The researcher asked editors to draw their process for filling knowledge gaps using an 8x11 sheet of paper or digital drawing program. Editors were instructed to label each part of their diagram and to provide a brief description so that others could interpret their process. Editors were then asked where an automated system would fit into this process and how it could augment their current workflow. These sketches and descriptions were uploaded to the online community, where other editors provided feedback and comments.

3.4. Participants

We recruited 19 active Wikipedia editors. 15 of our 19 participants answered all 8 prompts over the four week period. 1 participant answered 7 prompts, 2 participants answered 5 prompts, and 1 participant answered 4 prompts. Prior ARC studies have recruited similar numbers of participants, ranging from 13 [**MacLeod et al., 2016**] to 28 [**Walker and DeVito, 2020**], with a dropout rate of about 10%. We recruited participants through two channels: Wikiprojects focused on identifying and adding missing articles and organizations focused on filling knowledge gaps. For the full list of targeted Wikiprojects and organizations, see Appendix A.

Researchers must adhere to community protocols while conducting research within Wikipedia's editor community, as outlined by Wikipedia's "Ethically researching Wikipedia"[8] and "Research recruitment"[9] pages. Before recruiting, we created a project page[10] on MetaWiki which contained information about the project and allowed editors to comment on the project's goals and methods. In conjunction with our Wikipedia recruitment effort we also contacted organizations such as Whose Knowledge and Black Lunch Table, though these requests went largely unanswered.

Because this work focuses on understanding inequitable knowledge representation, which is created in part by unbalanced editor demographics, we attempted to recruit participants from a variety of demographic groups. Though we did not receive enough responses to follow a standard stratified sampling approach, participants filled out a pre-study screener (see Appendix B) which allows us to understand potential biases in our participant population. Our participants were relatively evenly split between male and female (8 male, 10 female, 1 non-binary), and tended to be between 30 and 50 years old (6 participants 30 to 39; 8, 40 to 49; 3, 50 to 59, and 2 over 60). Only 2 participants identified as Asian, while 15 identified as White and the remaining 2 abstained, but despite the lack of racial diversity we recruited participants from 14 different countries. Participants' education levels skewed predictably towards advanced degrees, where 9 of our participants had received a doctorate and a further 6 had received a masters degree. Likewise, the majority of participants worked in research or technology related fields.

[8] https://en.wikipedia.org/wiki/Wikipedia:Ethically_researching_Wikipedia
[9] ttps://en.wikipedia.org/wiki/Wikipedia:Research_recruitment
[10] https://meta.wikimedia.org/wiki/Research:Characterizing_Existing_Practices_for_Identifying_and_Mitigating_Knowledge_Gaps

We required that all participants speak English fluently and have significant experience creating and editing Wikipedia content. Though participants did not necessarily need to participate in projects or organizations that focus on filling knowledge gaps, all participants who responded to our initial screener engaged to some degree with these organizations.

3.5. Analytical Approach

We used an inductive approach adapted from Braun and Clarke's thematic analysis [**Braun and Clarke, 2006**] to develop cross cutting themes from this study. Because researchers have not yet investigated Wikipedia editors' strategies for identifying and filling knowledge gaps, this study did not rely on prior theoretical work as a basis for analysis. Instead we took a bottom-up, inductive, approach, where codes and themes were developed from reading and interpretation of the data.

Data analysis started with the following four guiding questions:

> ***RQ 1:*** *How do editors conceptualize and define knowledge gaps?*
>
> ***RQ 2:*** *What processes do editors currently use to identify knowledge gaps in Wikipedia?*
>
> ***RQ 3:*** *What processes do editors currently use to triage knowledge gaps in Wikipedia?*
>
> ***RQ 4:*** *How can socio-technical systems better facilitate existing knowledge gap identification and mitigation processes?*

Given that codes and themes were generated inductively, these guiding questions were treated as instructive and not exclusive. The specific process used to develop themes from the ARC data is outlined below.

To prepare for analysis, we uploaded all responses and comments to the coding software MaxQDA. Each "document" consisted of a response to a prompt or a comment, and we maintained document order to show conversation threading. Because all responses and comments were submitted as typed text by participants to an online discussion forum, transcription was not necessary.

Analysis for this project generally followed the steps outlined by Braun and Clarke:

(1) **Initial reading and memoing:** We conducted an initial reading of all of the content collected. We recorded high level ideas and notes that we developed during the initial reading, as well as possible codes to be used in the next step. The purpose of this step was to understand the general "shape" of the dataset without attempting to apply formal codes or structure.

(2) **Coding:** For each guiding question outlined above, we took two coding passes through the data. During the initial pass, we marked responses and comments that were germane to the specific question, and recorded a brief memo summarizing the text. During the second pass we then compiled each memo into one or more specific codes. When the memo did not fit an existing code, we created a new code. Throughout the coding process we reviewed existing codes to ensure that they did not overlap. If two codes overlapped, we collapsed them into a single code or added specificity to each. We compiled the final codes for each guiding question into a codebook, which is available in Appendix D.

(3) **Thematic Development:** Next, we organized codes into thematic categories using affinity diagramming. Themes included both examples and counterexamples, where participants might disagree with one another about a particular point or opinion, so long as both viewpoints are relevant to the overall theme. After affinity diagramming we developed labels and descriptions for each category. We assessed the categorized, coded data to ensure that themes were distinct from one another, and that the data supported each theme. If (and when) it was difficult to label and describe a category, we iterated on groupings to improve cohesiveness. Similarly, we continued to improve groupings until all codes fit into a thematic group.

(4) **Writing:** We developed a narrative around each theme by combining the descriptions developed in the previous step and specific quotes pulled from participant responses. We started by creating a rough outline from the thematic groupings and raw codes. We then returned to the data, adding participant quotes representative of each code or theme. Finally we wrote around each code to make sure we grounded each narrative in our data.

3.6. Results

During our thematic development process, three top-level groupings emerged: 1) editor conceptualizations of knowledge gaps, 2) potential causes of knowledge gaps, and 3) processes and workflows for mitigating knowledge gaps. In the following section we

explore each top-level grouping independently. In our discussion we then explore the implications of these groupings when taken together for the broader goal of working towards equitable knowledge representation.

We asked Wikipedia editors to describe scenarios in which they identified and worked with 1) missing content, and 2) biased content. After discussing missing and biased content extensively, editors then broadly defined the term "knowledge gap" and provided specific examples. As such, most knowledge gap definitions were influenced by prior responses and conversations about missing content and biased content.

3.6.1. How do editors conceptualize knowledge gaps?

Editors define knowledge gaps in terms of content that exists in other sources but does not exist in Wikipedia. For instance, knowledge gaps could be newly published content or current events that have not yet been added to Wikipedia. Conversely, knowledge gaps could be outdated content that needs to be updated. Knowledge gaps could also be content that exists in one language edition but not another.

> I would define a knowledge gap as knowledge that is not currently represented in traditional sources, such as books, scholarly journals, newspapers, even other multimedia such as radio, television, specialized trustworthy websites. I would also broaden the definition to knowledge that is not currently on Wikipedia, but it IS represented in some of those sources, but that nobody has introduced to Wikipedia yet; and to knowledge that is present in a particular language version of Wikipedia, but not in all. – p14

This perspective is in line with Wikipedia's "no original research" policy–e.g. Wikipedia can only reflect information and knowledge already available in other sources. Wikipedia does not attempt to create new knowledge or synthesize existing information, but rather acts as a tertiary reference to secondary sources.[11]

The scope debate (e.g. inclusionist vs exclusionist, quantity vs quality) extends to the definition of knowledge gaps. One continuous source of tension within the editor community focuses on defining what belongs in Wikipedia and what should be excluded [**Lam and Riedl, 2009, Mayfield and Black, 2019**]. Editors agree that Wikipedia should not contain all types of content; if Wikipedia's scope included misinformation, disinformation, unsubstantiated information, and fringe theories it would no longer be a reputable source. However, where editors define these boundaries differs across the editor community, which in turn affects how editors conceptualize knowledge gaps.

Broadly, one way editors define knowledge gaps is as missing or incomplete information, or omission of one or more narratives from a topic. This is a *quantity* perspective; knowledge gaps are defined by what does not yet exist but should exist.

> In Wikipedia I would define a knowledge gap as a set of topics or articles are missing due to one reason or another. As an example, I would classify villages in North Macedonia as consisting of a knowledge gap, as while some may exist as articles on Wikipedia, many don't, and many that do are not full fledged. In comparison, articles of villages in Australia would be in abundance and have more content in them. – p13

[11]See `https://en.wikipedia.org/wiki/Wikipedia:No_original_research` for Wikipedia's no original research (WP:NOR) policy.

> Missing content is when no content exists at all, and biased content is when some content exists, but the content does not encompass all perspectives around the subject. Biased content involves presenting one or a few (dominant) perspectives on a Wikipedia article, and not presenting other perspectives either deliberately or accidentally. – p17

The quantity based definition of knowledge gaps can be extended beyond articles to topic areas. For instance, some topic areas may be missing relevant information, creating a knowledge gap. In these cases, missing content is a type of bias, where Wikipedia may be biased towards including information about a one topic while also biased against including information about a different topic.

> Recently I've been looking at the history of Glasgow and the slave trade, and have found lots of articles about tobacco lords which don't mention the slave trade, or where they got their money from. In that case it's an omission of content which leads to a bias in the way that those people are presented. – p8

The *quality* perspective defines knowledge gaps as content that does exist, but should not be contained in Wikipedia because it is incorrect, distorted, biased, or not neutral in its point of view. These types of definitions were usually framed around biased content. Knowledge gaps may include content that is not backed up by sources or by content that supports a point of view that is not held by the majority. Knowledge gaps may also include content that is intentionally or unintentionally miscatagorized, misgendered, or misnamed such that the content does not accurately represent the subject matter.

> Biased content is where some factual information might be presented, but it's presented uncritically, without a consideration of how the source itself might be biased, or how a narrow range of sources might not present the entire picture. In other cases, biased material might soften unpleasant or uncomfortable aspects of history. This differs from missing content, which simply isn't on the encyclopedia. Biased content is there, but doesn't capture the whole story, and is written to (consciously or unconsciously) downplay or omit some facts. – p0

Exclusionist knowledge gaps can also apply to specific articles or entire topical areas. For instance, a single person or entire cultural groups of people can be misnamed or miscategorized. Non-neutral point of view content can apply to a single article, or it can create systematic bias throughout an entire topic area.

However, bias is subjective, which creates ambiguity when deciding which narratives should be included and which should not. Editors acknowledge that not all knowledge gaps are created equal. For instance, editors prioritize certain underrepresented groups (e.g. women and other minorities), but they deprioritize other knowledge gaps that nevertheless encompass legitimate, notable content.

> The entire platform runs on constant vigilance...but I suppose I am thinking [specifically] about the...effort needed to 'protect' certain types of content over other types of content. – p11

> Getting closer to [achieving] the sum of all knowledge is not necessarily the biggest problem...The sum of all knowledge is an achievable goal,

> and Wikipedia will reach [that goal, so we should] focus more explicitly on the diversity of knowledge whilst we get there. – p9

> I was putting together a list of potential articles & topic areas pertaining to Scotland's slave trade history (there's quite a gap) and found myself with a list of tobacco lords. Some of their articles don't mention the slave trade at all, some simply don't have articles. And of the latter, I found myself wondering if I even wanted to direct my energy toward writing their histories. There are other, untold stories, that I could spend my time on. – p8

3.6.2. What do editors think causes knowledge gaps?

The following section outlines editors' ideas concerning the causes of knowledge gaps. These fall in several broad categories: the contributor gap, point of view (POV) editing, and Wikipedia policy.

3.6.2.1. The Contributor Gap. Wikipedia's contributor gap has been outlined extensively in prior work [**Redi et al., 2021**], but it appears consistently throughout participant responses and is therefore worth highlighting. Simply put, Wikipedia's editors do not have enough time to add, improve, and police the quantity of high quality content that Wikipedia aspires to include. Editors in our study rarely struggle to identify content that needs to be created or improved, but rather cannot keep up with and triage their existing to-do lists. The editor community needs more person-hours to maintain and add

content, whether through increasing the number of contributors or increasing the number of hours existing editors can contribute [**Geiger and Halfaker, 2013**].[12]

Participants in our study did note that not all hours contributed are created equal. Learning to create quality content while navigating Wikipedia's complex policies and politics takes time and experience. While newcomers are important for the health of the community, low quality content creates additional work for experienced editors.

Participants also mirrored prior work [**Hecht and Gergle, 2009, Hecht and Gergle, 2010b**], noting that demographic asymmetries in Wikipedia's editor population can lead to more systematic knowledge gaps. Participants in this study explain that editors tend to both work on content in which they are subject matter experts and content that personally interests them. This makes intuitive sense; Wikipedia's editor community is entirely volunteer, and there is little incentive for editors to work on topics outside of their personal interests. Furthermore, creating high quality content requires an immense amount of knowledge about the subject, which discourages editors from working outside their areas of personal expertise [**Hecht and Gergle, 2009**].

> There are systemic content deficiencies at WP. Maybe that's a way to phrase it. Since content people tend to gravitate towards what interests them, they move away from things that do not. This can be cultural; people tend to write about what they know and do not write about what they do not know. But it can also be as simple as not wanting to write about boring things. For example: there is a systemic content deficiency related to 19th century African and Asian political biography,

[12]See also: `https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_a_work_in_progress`

I presume. Some of the reason for this is cultural; some of this is related to lack of familiarity [or] of lack of access to sources...There is also a systemic content deficiency related to American business companies and executives. The content that exists tends to be putridly bad. Much does not exist at all, or exists in such a rudimentary form as to be useless. This can not be blamed on cultural unfamiliarity. It's just really boring. Far more entertaining to write about a TV star or a sports team or a new bottled drink or what have you. - p15

While I will write articles about female Australia artists when I am supporting such initiatives (as I currently am with our national gallery), I only do it while I am at those events (or in the immediate aftermath of them while I dot the "i"s and cross the "t"s). Art simply isn't my interest and, having "done my bit" for art, I revert back to my usual history and geography focus in my own state (with no shortage of missing topics and content). It doesn't matter how huge the gap of missing Peruvian insects or Nigerian marathon runners is on Wikipedia, *I* am not going to fill it and I don't think I am atypical. - p4

Part of it is that people write about what is interesting to them. There is an interest in things that comes from locality. The number of very active Wikipedians from London or San Francisco is huge, and the press from those places is equally gargantuan and varied, so urban topics get picked off rapidly and the output on these things is cumulative. However there are apt to be zero very active Wikipedians from Corvallis, Montana, for

> example, so if there were (making this up) a historic building there, the pool of people apt to write on that is small. - p15

Because editors tend to work on what they know, demographic asymmetries in Wikipedia's contributor base lead to asymmetries in the types of content that editors add and improve [**Hecht and Gergle, 2009**]. Prior work shows that Wikipedia's editors skew white and male, and they live in traditionally "western" cultures. The reasons why the editor population has remained relatively homogeneous are numerous and complex, but both prior research [**Bryant et al., 2005**] and participants in this study suggest that various gatekeeping practices lead to low newcomer retention, especially among underrepresented populations. Outright hostility and bias–both conscious and unconscious–drive away newcomers, particularly those newcomers who would add diversity to the community's collective perspectives. Furthermore, Wikipedias' complex rules and policies favor editors already educated in western research and publication techniques. In order to contribute, editors must also have access to technology and leisure time to contribute. Put together, these factors help maintain a relatively homogeneous contributor base which results in less diversity of content.

3.6.2.2. Point of view editing. Point of View (or POV) editing represents a second source of knowledge gaps in Wikipedia. POV editing usually occurs when editors create content not supported by reliable sources in order to support a specific narrative, though in some cases POV editing may include the removal of established and well-sourced facts. Generally this behavior results in content that should not exist, either because the subject is covered inappropriately or because the subject does not meet notability requirements.[13]

[13]For a discussion of Wikipedia's policy on point of view editing (WP:POV), see: `https://en.wikipedia.org/wiki/Wikipedia:NPOV_dispute#POV_pushing`.

POV content results from both conscious and subconscious biases. In more intentional cases, editors may cherry-pick or obscure sources in order to support biased narratives or to bolster the credibility of a pet theory, personal acquaintance, favorite product, or other non-notable subject matter. Participants also noted more harmful examples of POV editing in which editors use similar tactics to promote racist and sexist theories about underrepresented groups, or to spread outright misinformation about politicized topics.

> We have a case such as this on [Portuguese Wikipedia] where the biography of Aristides de Sousa Mendes, a portuguese diplomat during the 2nd world war that issued visas to people trying to escape Germany against the orders from our own dictator at the time, where an editor continuously cherrypicks facts to completely skew the POV towards telling a completely different story than the majority of scholars portrays. The Wikipedia biography has been denounced by historians several time [sic], most recently this year,[14] but it is extremely difficult to balance the POV of the text, because the editor is extremely skilled in walking the fine line between policies, always cites his sources and never straight out reverts any changes. He bides his time and gradually returns the article to say what he wants it to say. – p14

However, POV editing also results from subconscious cultural biases that authors may not know to address. For example, participants mentioned that sexist and racist content is often perpetuated by outdated mainstream sources and is reproduced by editors more out of ignorance than maliciousness.

[14] https://www.publico.pt/2020/06/21/politica/noticia/versao-falseada-aristides-sousa-mendes-wikipedia-1921080

Editors in our study indicated that POV content could appear anywhere, but that certain topics are more likely to attract POV editing. Common politicized topics such as vaccine safety, abortion, and LGBTQ issues tend to attract POV editing, though baised content also appears in less obvious areas, such as Soviet history and horse racing.

3.6.2.3. Policy. Wikipedia attempts to define the encyclopedia's scope through layers of policy.[15] In theory, policy allows editors to create uniform guidelines that determine what content constitutes encyclopedic knowledge worthy of inclusion, and what content does not meet it's qualifications. Ideally editors could leverage policy across a wide array of topics to somewhat objectively and uniformly determine whether content belongs.

The core of Wikipedia's inclusion criteria rests on the concept of notability, which in turn rests predominantly on reliable sourcing. Notability determines which concepts are important enough to be included in Wikipedia; whether a topic is notable in turn depends on whether it has been covered by reliable secondary sources.

3.6.2.4. WP:NOTABILITY. While notability and sourcing guidelines aim to provide overarching criteria that allow editors to determine what content to include, participants in our study noted that interpretation of both policies can be highly subjective in a large number of topic areas. Some topics–such as national monuments, heritage sites, and parks–map well to official lists that confer both notability and completeness. These official lists help editors determine what subjects merit articles and how comprehensively Wikipedia covers that topical area. As one participant explained:

> Completeness projects are most feasible when there is a real-world list that corresponds to Wikipedia notability e.g. of "worthy of gazettal"

[15]Shorthand for Wikipedia's policies follows the form: WP:<name-of-policy>. For instance, editors refer to Wikipedia's no original research policy as WP:NOR.

> (place names), "worthy of heritage listing" (heritage sites), "member of parliament" (person), etc. Where there are clear-cut notability rules in Wikipedia that relate to well-defined lists, completeness can be determined – p4

However, many topics are not suited to official lists or strict inclusion criteria. For less clear cut topics, inclusion criteria may unintentionally reinforce biases written into the notability guidelines themselves.

> If we work with something...like "all notable artists", then if you look at the notability rules for "creatives", there are a number of criteria, some of which are hard to demonstrate. The one that is probably most used relates to "significant" exhibitions and works in the collection of "major" galleries. But not all galleries provide a public list of all of their holdings and of the works included in all of their exhibitions. So the notability of an artist in terms of "should have a Wikipedia article" is now dependent on their works being held by galleries with a public records of catalogues/exhibitions. – p4

Furthermore, non-specific inclusion criteria will always be interpreted and enforced through the lens of editors' biases. As the author of the above example noted, whether a gallery is "major" or whether an exhibition "significant" is vague and therefore open to interpretation. For example, a major street artist would not display work in a gallery, which would therefore define that artist as not notable.

Interpretation of Wikipedia's notability guidelines does not affect all topics and types of knowledge uniformly, which in turn creates systematic knowledge gaps. Due to the

relatively homogeneous editor population, editors may judge certain topics to be inherently more notable while other topics are systematically excluded. For instance, multiple editors in this study noted that the notability of well documented female historical figures is questioned far more frequently than similar male counterparts [**Wagner et al., 2016, Wagner et al., 2015**].

> One of the issues with wp:notability...is that the encyclopedia by necessity recognises only certain types of knowledge. In the past I've come across a slew of issues surrounding notability pertaining to women's history, where social systemic bias has marginalised that history and those individuals, and wp:notability requirements re-entrench that bias. – p8

> The notability standards and reliable sources requirements were intended as a shield (to prevent articles about crackpot theories and garage bands) but have now been used as a sword... in some cases to perpetuate systemic bias, or even just unconscious, unexamined bias. – p10

3.6.2.5. WP:RELIABILITY and WP:VERIFIABILITY. Wikipedia's reliable sourcing policies aim to provide more objective guidelines that editors can leverage to determine notability. If a subject has been appropriately covered by a reliable source, then that subject meets Wikipedia's notability guidelines. As stated in WP:Notability:

> A topic is presumed to be suitable for a stand-alone article or list when it has received significant coverage in reliable sources that are independent of the subject.[16]

[16] `https://en.wikipedia.org/wiki/Wikipedia:Notability`

Editors in our study highlighted general lack of sources and difficulty in locating sources as a major challenge for writing articles. Although Wikipedia relies on verifiability to indicate notability, a lack of available sources does not necessarily indicate a lack of notability. For a variety of reasons, academic and news sources may not cover a specific subject, or those sources may be hard to locate [**Schneider et al., 2012**]. Editors identified skirting paywalls that limit access to scholarly research and locating expensive out of print publications as major barriers to article creation.

Furthermore, participants in our study noted that reliable sourcing guidelines systematically exclude specific types of information due to subjective interpretations of reliability. Just as nobility determinations frequently reflect the interests and cultural backgrounds of the editor population, reliability determinations frequently reflect those editors' training in the Western academic research and scholarship. While this style of research does generally produce high quality, verifiable, and reproducible knowledge, it also prioritizes written, English language, peer reviewed source material.

Unfortunately, certain types of knowledge tend not to be documented in peer reviewed, English language journals. Multiple editors in our study worked with and documented cultures that maintain strong oral histories, which scholarly publications–and therefore Wikipedia–have no way of citing [**Gallert and Van der Velden, 2013**]. As a result, histories of entire populations not documented in peer reviewed sources are excluded.

> The example I'll talk about here though is of The Tinkers' Heart of Argyll, the only permanent monument to the travelling community of Scotland. I was working in the museums sector, and came across a news article about the site and the campaign to have it listed - listing

> monuments relies largely on written sources to prove notability, and because the proof of its significance lies mostly within the oral history of the community, the application had initially been declined. I'd looked it up on Wikipedia to find out more about it, and there wasn't an article. I created a list of sources in a spare sandbox... and then it stayed there for ages. The travelling community doesn't leave many written sources - see the aforementioned comment about oral history - and some of the sources that I found online weren't the best, a lot of self-published blogs about the campaign. What did help was that a national newspaper had picked up the campaign, so I could use those as sources, and there had been a statement from the body who lists monuments in Scotland, providing more proof as needed on-wiki for notability. – p8

Wikipedia does not require sources to be written in the language edition's primary language, but English Wikipedia editors noted that prioritising English language sources deprives the encyclopedia of specific types of content. Editors in English Wikipedia tend to see English language sources as more reliable–or at least easier to verify–which creates a bias towards content covered by English language sources [**Ford et al., 2013**].

> I am sure many countries are under-represented in en.WP because their sources aren't in English, and their citizens don't generally have good English language skills, and their own language is not widely spoken outside of their own country. Again, it comes to the link between content gap and contributor gap. – p4

While Wikipedia's sourcing guidelines systematically exclude certain types of information, they also include sources that promote one-sided and outdated points of view. Participants in our study frequently encountered Point of View articles about underrepresented groups that were supported by outdated sources considered reputable at their time of publication. Due to Wikipedia's reliance on peer reviewed publications, these pseudo-reliable sources advance Western colonial and imperialist narratives.

> [I've had] similar problems with Native American topics, where the "reliable sources" were written by colonizers and by modern standards are quite racist, distorted, etc. I was floored to actually have a major editing debate once over using an 18-century source that was absolutely horrible, but someone argued that it was somehow an "official" source, so reliable (it said horrible things about the Native nation of discussion) – p10

> Sources in public domain (and we historically used them a lot on Wikipedia) are old and with an old fashion perspective on the world. which means for example that they present ethnic groups based on a XIX century vision of the world (times in which we used to measure skulls). Also we use sources we know and often wikipedia contributors on certain topics they studied in high school (and not in university) have old references (in particular they miss a lot of post-colonial perspectives). – p6

> I was led to a powerful example of this by a grad student colleague, who pointed me to an article about an 18th century colonial military service member who was also an author. The more biased version of

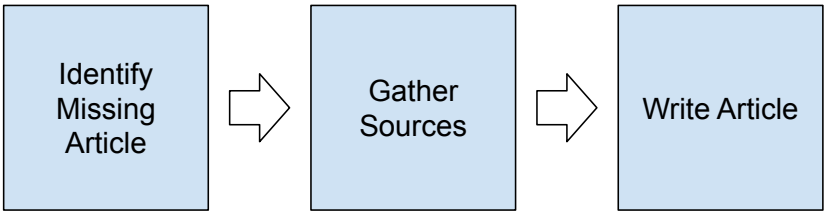

Figure 3.1. A flowchart depicting editors' generalized workflow for creating new articles.

> the article takes the primary source – a famous narrative written by this person – as entirely factual. Which shows a deep misunderstanding of the reliability of narratives by 18th century white male soldiers in the Caribbean. In particular, the more biased version of the article uses specific instances from the narrative to characterize the writer's relationship with an enslaved woman as romantic. – p0

3.6.3. How do editors identify Knowledge Gaps?

Editors in our study created content related to a wide range of different topics across several different language editions. On a surface level, editors generally followed the process outlined in Figure 3.1 to create a new article, first identifying an unmet informational need, gathering sources, and finally writing the article text.

Figure 3.1 is intentionally oversimplified. Each step of the knowledge gap identification and content creation process is complex and nuanced, and depending on the context of the subject and the particular task, each editor leveraged one or more unique workflows. In this section we focus on the various methods editors employed to identify missing articles.

While no two editors used exactly the same method, workflows for identifying knowledge gaps tended to fall under two categories: *Serendipitous Discovery* and *Systematic Search*.

3.6.3.1. Serendipitous Discovery. Some editors in our study stressed that they do not have systematic methods for identifying missing content. Instead, these editors tend to serendipitously "stumble upon" information missing from Wikipedia in their daily routines. As one editor noted when asked to reflect on this study as a whole:

> I gently criticize [the approach of this study] for looking at editing in an excessively formalized manner. There have been a number of responses that described not using methodical or systematic methods but just doing as we see fit at the moment (which will be different at different times). Sure, I understand that for a scholarly study, you need to observe patterns, generalize and draw inferences about them. And I think that approach distorts many editors' approach to editing which is often unsystematic and based on a combination [of] serendipitous encounters with Wikipedia and the world at large. – p7

The most basic method by which editors serendipitously identify missing content is through reading Wikipedia, or *Active Reading*.[17] When reading an article, editors notice incomplete content and fill in the missing information.

> I was looking at the Constructive journalism page. It mentioned the first dissertation on the topic, including the author and institution. I

[17] *Active Reading* typically means "[A] deep, focused reading process...Active reading can be characterized by the greater demand it places on the reader and her media and tools. Active reading (AR) frequently involves searching, highlighting, comparison, non-sequential navigation, and the like." [**Tashman and Edwards, 2011**] While the process described by Tashman & Edwards does not exactly map to the process described by our participants, we believe the general concept still applies.

> wanted to know what year it came out, but the article didn't say. I did a little bit of web searching, and found the date, and then added it to the article. – p5

> As an alumnus of Oregon State University and resident of the town in which it is located, I have a local interest in the history of the school. The main article on the school's men's basketball program, just as an example, has a footer template listing individual season histories. Many of these are red links – there is no extant article on the 1902-03 season, for instance...So, if I get a wild idea to kill four or five hours doing some light research on something OSU sportsy, I might dive into something like that. Is that an urgent, lacking article? No, of course not. The point would be to find an unplowed field and to have a little fun digging up what can be dug from newspapers and, if I have one around, yearbooks. – p15

As illustrated by the above examples, incomplete information could be as simple as a missing date or as extensive as an entire non-existent article. While this method seems straightforward and obvious, it illustrates the unintentionality that frequently characterises knowledge gap identification. Some editors find knowledge gaps not through comprehensive or thorough search, but through coincidental encounters while completing other tasks.

Failed Search represents a second way in which editors serendipitously identify missing information. While conducting other tasks–for instance writing or conducting research–editors "look [for information on Wikipedia] and fail to find something (p15)". One participant in our study provided the following example:

> There was an early activist in the early 20th century American radical movement named Eadmon MacAlpine. If I were merrily writing along and in need of some detail on his life, I'd be apt to jump over to WP and run a quick search. Nothing there? Oh, I need to do that one. So I will create a red link on my user page as an aid to memory and move along with my day, and then when time and motivation allows, circle back and write the piece. – p15

The differences between serendipitous knowledge gap discovery through active reading and through failed search are subtle. Like active reading, failed searches are not systematic; a failed search occurs unintentionally and without the objective of identifying missing content. Yet unlike active reading, failed searches indicate a real informational need Wikipedia was unable to fill.

Cross Referencing external sources represents a third way in which editors identify missing information. Many editors are subject matter experts who conduct research professionally, or at least external to Wikipedia. When reading new material about subjects that fall within their professional areas of expertise, editors tend to know which subjects are already covered in Wikipedia, and conversely where knowledge gaps likely exist:

> As a librarian and trained musicologist, I have a pretty good knowledge of my field and nearly always think of topics not covered in Wikipedia.

> I do not approach editing in general or knowledge gaps in a systematic fashion. Rather, I come across topics that attract my interest and then elaborate, expand or create content based on knowledge gaps. – p7

Many editors cross reference external sources with Wikipedia, checking whether Wikipedia covers the new or obscure information. In these cases, examples of external sources include newly published scientific information, obscure or newly published content about historical figures and events, or sufficiently notable current events.

> I have a habit now of searching Wikipedia for people when I learn about them. Currently my own work and research is focused on various social justice movements so I'm not having any trouble finding content gaps! When I see a gap or incomplete article, I open up a spreadsheet I created, add the name and start dropping in media article links for reading and possible citation. – p11

> I suppose, in thinking about it, that there is a third form of missing information: details about newly published books or arcane names to be added to lists. One knows when they bump across it that this material is almost certainly missing and that it can be quickly and easily added – it isn't something that one needs to set aside half a day to do. Yesterday when I was scanning up microfilm I spotted the name of a newly elected Socialist mayor of Granite City, Illinois. I knew of the page "List of elected socialist mayors in the United States" and knew there was no more than a 10% chance that the name was on there. Bing-bang-bong, done. – p15

> For a primary example (which I mentioned in my earlier post) is when I did my PhD I read about a lot of topics connected to my field of research and when I saw an article did not exist for a thing in relation to the topic (radiologically isolated syndrome) I decided to create it using the sources I knew of outside of Wikipedia. – p13

Active Reading, Failed Search, and Cross Referencing exemplify serendipitous, non-systematic methods of knowledge gap identification. Critically, these methods differ from others discussed later in this chapter because, in these examples, editors do not begin with the goal of identifying or filling a knowledge gap but rather serendipitously encounter missing information and subsequently add it to the encyclopedia. The result of serendipitous discovery is that editors add missing content based on a variety of factors–e.g. topics the editor happens to be reading or information the editor needs on a particular day–that may reinforce existing biases towards certain types of content. For instance, a white, male, editor from the United States or Europe may be more likely to read articles about western military history, which would in turn mean that the same editor would be more likely to serendipitously encounter missing content about military history through active reading. At a large enough scale, demographic asymmetries could unintentionally result in certain topical categories receiving more attention than others.

3.6.3.2. Systematic Search and Curation. While some editors emphasized the somewhat unintentional nature of their editing process, others leveraged more systematic approaches. In our study, editors' systematic approaches almost always started by first identifying a topic space and subsequently defining which subjects should be covered within that topic space. For each subject, editors then follow the general process outlined

in Figure 3.1, gathering sources and creating content. This general approach to systematic knowledge gap identification makes intuitive sense. Knowledge gaps are defined by their relationship to the sum total of content that should exist, so editors must first define "completeness" for a specific topic before they can systematically create or improve content to reach complete coverage.

Editors in this study primarily defined completeness through external lists generated by subject matter experts. As noted earlier, external lists help "confer notability (p18)", which in turn defines the scope of a topic area. In cases where external lists already exist, editors create "hit lists", "red link lists", or "completeness lists" through a combination of manual and programmatic methods.

> There are two main types of "workflows" I use when identifying missing content. I either have a list of topics that I want to improve coverage on Wikipedia, such as biological taxa or cultural and natural heritage, and I follow an alphabetical list to see what content is missing. In the case of biological taxa I have tried to automate this process several times using a basic template article where I fill in the specific gaps and using software to batch create the articles. The results are mixed in my view, as I forego the whole making the infobox hassle this way, but the resulting articles are short and insipid. – p14

> I compiled a list of towns, suburbs and localities from the Queensland Place Names database (the official gazetteer of Queensland) which is available as a spreadsheet. I compiled a similar list of Wikipedia articles by using Petscan based on various likely categories using its export to

> spreadsheet tool. I then compare these to spot the differences. This reveals a lot of things including most immediately missing topics from Wikipedia. It also reveals miscategorisation of Wikipedia articles (e.g. Wikipedia articles categorised as towns but which are not gazetted as towns). More tediously, one can check the lede para and infobox of each article to confirm that they describe the place as the correct type of place and contain a citation to the official place name index. The notability criteria for "place" articles in Wikipedia is being officially gazetted, so every place article in Wikipedia should contain a citation to wherever it is officially gazetted (many don't). – p4
>
> Over the years I have also cross referenced articles from resources such as the Dictionary of Irish Biography (DIB) to see who was missing (generally focusing on women)...Often I will work from a text or corpus which tends to confer notability, like the DIB, or the database of National Monuments, or the databases from the National Parks and Wildlife Service, and work out from there. - p18

For many topics, external completeness lists do not already exist. In these cases Wikipedia editors work with subject matter experts to create lists of notable subjects. While these lists do not automatically confer notability to the rest of Wikipedia's editor community, experts can curate source material and citations to accompany completeness lists.

> I'm very lucky in that I connected early with local historians who already had a good sense of where history could be expanded. For example, with

> classroom projects, I've first talked to our archivists who steward a wonderful collection on underrepresented groups in Boston. My stance is that I know that there are certain groups who are generally underrepresented in history/perspectives, so I'm starting from that as a base. Our archivists helped come up with a list of people and events that were (a) important in Boston's history and (b) had a relatively accessible group of secondary sources. This was all very time-intensive, basically done by hand: they give me a list, and I look for which articles would be easier for new editors (i.e. the students I work with) to handle. – p0

> A lot of the time, the missing content I work on has been identified by existing groups - Women in Red, or a group with whom I'm doing an editathon. In those cases we're drawing either on Wikidata redlink lists, or the expert knowledge of a group who are familiar with a topic area, and have done a content audit to identify gaps, so in that case I'm relying on the knowledge of others to identify a gap which I've helped fill. – p8

In the above examples, editors use completeness lists to define the scope of a particular topic and work to complete all articles within that defined scope. However in other examples, once a topics' broad scope has been defined, editors actively deprioritize certain subjects in order to focus on known underrepresented subsets of a topic area. In these cases, editors further refine a topic area's scope based on known biases. For instance, several editors in our study focused specifically on creating content about women or people of color while actively avoiding creating articles about men.

> Lately my interest has gravitated towards the lack of women in the field of music and I create articles when I know that the topic can help redress the gender bias issue. I've created or contributed a few articles on women who can not be found in reference sources, so it can require a great deal of research. – p7

> Being involved in Women in Red, I do try and keep my new articles on men to a minimum, but it's hard! There are a lot of important Irish men still missing. At the start of my 366wikidays I was writing 1 male biography for every 19 women, but I have been writing about places more recently, and living black and people of colour in Ireland. – p18

> Re: editing on men - I was putting together a list of potential articles & topic areas pertaining to Scotland's slave trade history (there's quite a gap) and found myself with a list of tobacco lords. Some of their articles don't mention the slave trade at all, some simply don't have articles. And of the latter, I found myself wondering if I even wanted to direct my energy toward writing their histories. There are other, untold stories, that I could spend my time on. – p8

3.6.3.3. Distributed Systematic Curation through Internal Markers. Working from external completeness lists is not feasible for every topic covered by Wikipedia. Publicly available, up to date, comprehensive lists that are produced by reliable sources do not currently exist for all topics. While completeness lists can be created by subject matter experts for a particular topic, this requires large time investments from a relatively limited group of contributors. Given both the breadth and the volunteer ethos of Wikipedia,

creating and maintaining external completeness lists for every topic would be impractical if not impossible.

In order to distribute the work required to create and maintain completion lists, the Wikipedia editor community maintains an ad hoc socio-technical system that produces internal completion lists within a topic space. This system relies on two main technical elements: 1) light-weight markers that indicate missing, incomplete, or low quality content, and 2) aggregation systems that create lists of these markers.

3.6.3.4. Red Links as Markers. Internal markers take multiple forms, but Wikipedia's most common marker is the red link. Red links indicate that "the linked-to page does not exist—it either never existed, or previously existed but has been deleted."[18] Linking to a non-existent page in some ways negates the intended purpose of a hyperlink; links traditionally indicate "a reference to data that the user can follow by clicking or tapping,"[19] but red links reference no data and cannot be followed. However, editors create red links as a light-weight marker of missing content, or data that should exist but does not. Creating a red link requires a minimal time investment compared to writing a short article, but the red link serves as an indicator or a reminder that the article should be written. Importantly, red links can exist anywhere on Wikipedia. Editors commonly add red links to the text of articles or as items in lists, as well as to-do items on user pages and wikiproject pages.

Although red links were most referenced in this study, editors mentioned a variety of other internal markers. For instance, stub articles serve a similar purpose as red links, but slightly farther along the production pipeline. Stub articles contain a minimal amount

[18] `https://en.wikipedia.org/wiki/Wikipedia:Red_link`
[19] `https://en.wikipedia.org/wiki/Hyperlink`

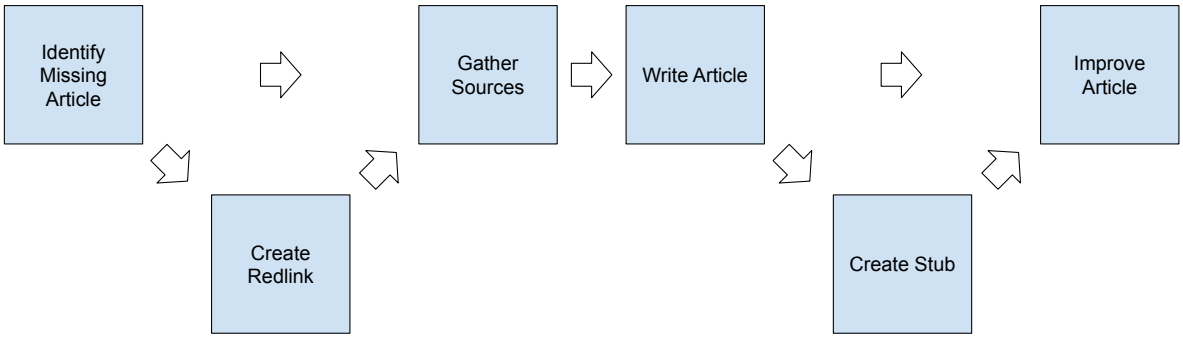

Figure 3.2. Generalized article creation flowchart that leverages markers. Editors can enter or exit the workflow at the "identify missing article", "create red link", or " create stub" steps.

of text and sources while indicating that the subject requires more content. Similarly, editors tag existing articles with a variety of cleanup templates (e.g. Template:Systemic bias,[20] Template:POV,[21] or Template:More citations needed[22]), indicating that the article needs improvement at some point in the future.

At a basic level, red links and other markers facilitate the redistribution of labor required to identify and fill knowledge gaps. Whereas in Figure 3.1 a single editor identifies an article that needs to be created, finds the relevant sources, and finally writes the article, Figure 3.2 illustrates a workflow that leverages markers to distribute this process. Editors can enter or exit the workflow at the "identify missing article", "create red link", or "create stub" steps. For instance, If an editor identifies a missing article but does not have the time or expertise to gather sources and write, that editor can create a red link and effectively pass the remaining labor to another editor.

[20] https://en.wikipedia.org/wiki/Template:Systemic_bias
[21] https://en.wikipedia.org/wiki/Template:POV
[22] https://en.wikipedia.org/wiki/Template:More_citations_needed

> Perhaps as an example, I could discuss Irene Reed's case. While not Yup'ik herself, she was one of the people responsible for so much of the work done on Central Yup'ik. I don't remember what I was looking for in the first place, but I found a massive repository of information, scans, PDFs, etc. about the community and its language. Amongst all of it was her obituary. When I tried to find out more info about her life, no Wikipedia article came up. As I often get sidetracked with other things, I put it up on the English Wikipedia's Women in Red talk page: [url][23]. (I do this a lot, because I have poor access to many of the sources other people have and because there are some phenomenal people in that project who can find information on anything you throw at them!) This led to some other people getting interested in it and they wrote a lovely article about her. My contributions were quite minor in the end. – p2

Multiple editors in this study indicated that stubs are easier to expand than to create. These editors would therefore focus on stub creation while leaving article improvement to less experienced editors. As an example of this workflow, a subject matter expert with limited time might stumble on a red link, identify enough sources to create a stub, and pass the labor of writing the article to another editor while moving on to create another stub.

> For those of us who are not scared, I say "make stubs" as much as you can because of we have plenty of people willing to expand stubs but not to create them. I am a great believer in the 2 sentence 2 citation

[23] https://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Women_in_Red/Archive_80#E._Irene_Reed_or_Irene_Reed

> stub. For the benefit of the non-prolific-stub-creators, the first sentence defines the topic. "Joe Bloggs (1900-1970) was a farmer and politician in Queensland, Australia". The second sentence establishes notability. "He was a Member of the Queensland Legislative Assembly from 1950-1960". Add two citations to support these two sentences. Done! – p4
>
> Even when I have a keystone notability reference, I'm often very anxious about "just" writing a stub about someone or something, and will pressure myself into holding off until I can dedicate more time or resources to a topic. That is a huge shame, as if I just publish the stub, another editor could come along and add the extra info they have ready access to rather than me having to expend lots of extra resources to find them. – p18

Not all editors use markers in their workflows. Some editors prefer to independently identify missing content, gather sources, and create the resulting article. In these instances collaboration only begins after the full article is published when other contributors add or edit content.

> When I'm passionate I drop all editing and try to work up an article. Otherwise I add to it when I have time. I'm not the kind of person that will publish a stub and work on it. Rather, I try to write a complete article in my sandbox, and when I feel I can't do anymore, I publish it. – p7

This comparison illustrates an important element of red links and other markers. While some editors use markers to distribute labor among multiple community members,

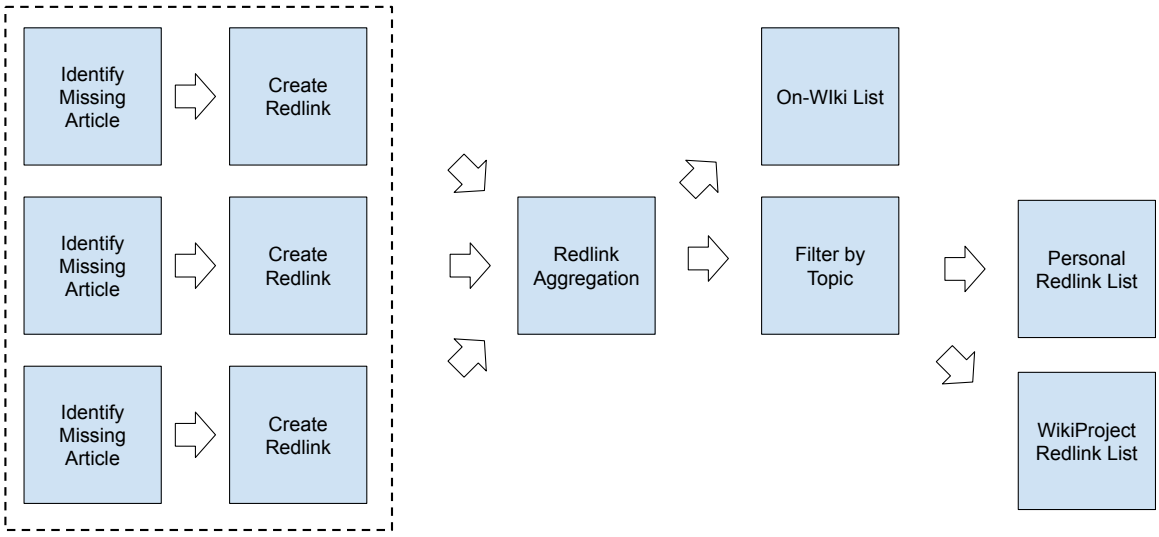

Figure 3.3. Generalized redlink aggregation for systematic knowledge gap discovery

editors can choose the degree to which they integrate markers into their workflows. Due to the diversity of editing styles used within Wikipedia's community, inflexible systems that force editors into specific workflows tend to receive push back and may ultimately fail.[24] Red links, stubs, templates, and other markers have been widely adopted by the community, but they do not require editors to create content in a specific way.

3.6.3.5. Aggregation Systems. Aggregation systems represent the second key component Wikipedia editors use to maintain internal completion lists. While red links are useful for indicating missing content, without an aggregation system they merely facilitate opportunistic editing practices. Aggregation systems allow editors to document the superset of all known missing content, systematically creating completion lists from internal markers. Creating such a superset then allows groups of editors to approach content

[24]For archives of editor commentary about the AFT, see: `https://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/Article_feedback` and `https://en.wikipedia.org/wiki/Wikipedia_talk:Article_Feedback_Tool/Version_5/Archive_1`.

creation through top-down systematic methods in addition to bottom-up opportunistic strategies.

The most basic form of aggregation used by Wikipedia editors is personal to-do lists. Many editors create lists of red links on their user pages as reminders of previously identified missing content. While time consuming to create and not comprehensive, personal to-do lists allow editors to choose which knowledge gaps most need attention rather than relying on opportunistic methods.

> I still rely a lot on old fashioned to do lists in my user pages. I started doing this early on in the work of Wikimedia Community Ireland, as we started doing editathons, we needed lists of articles with issues that new editors could work on. I would do this by systematically going through categories for relevant articles as well as articles marked as stubs by WikiProjects (usually WikiProject Ireland). – p18

Many editors collaborate on personal to-do lists with other editors who share common topical interests. This allows editors to distribute knowledge gap identification and content creation for a specific topic among several contributors.

> I know a number of editors who do very similar work to me, one of whom I actually will share my to do lists and red lists with and who knows she is free to take any from my lists to write up herself. – p18

> People do compile to-do lists and just leave them around for anyone interested, e.g. [this list][25] which is an approach that rarely achieves any

[25] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Australia/To-do

> kind of completeness but does at least direct people looking for a topic to one that someone perceives a need for. – p4
>
> ...for indigenous women or non-indigenous women who are important to the communities in the communities own opinion I do more systematic searches. This info I add to [this list].[26] This messy list started out as a list of mainly Saami women and as a to do list for myself, but it's been linked from other projects and people add in new people themselves, so now it's more a collaborative list than any solo project, which I love. – p2

Wikiprojects offer a more permanent social organization for creating and maintaining to-do lists about a specific topic. Wikipedia defines a Wikiproject as:

> *...a group of contributors who want to work together as a team to improve Wikipedia. These groups often focus on a specific topic area (for example, WikiProject Mathematics or WikiProject India), a specific part of the encyclopedia (for example, WikiProject Disambiguation), or a specific kind of task (for example, checking newly created pages).*[27]

In practice, a Wikiproject's contributors generally maintain lists of red links and articles that require improvement. Whereas a personal to-do list is built around the interests of a particular individual, a Wikiproject's to-do list is constructed around the mission of the organization. This leads not only to a theoretically more durable effort to complete to-do lists, but also a more centralized, systematic, and therefore comprehensive approach

[26] `https://en.wikipedia.org/wiki/User:Yupik/Redlinks/Indigenous_Women`
[27] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject`

towards building lists of red links or lists of articles that need improvement. Editors in this study referenced WikiProject Women in Red as by far the most prolific and systematized effort to quantify and rectify a knowledge gap, though many editors participated in a range of other Wikiprojects as well.

The centralized nature of Wikiprojects allows editors to participate in a systematized effort to address knowledge gaps without undertaking the requisite organizational overhead. Participants often mentioned using a Wikiproject's red link list as a method for choosing which content to produce.

> A classic everyday example is looking at the lists in Wikipedia:WikiProject Missing Encyclopedia Articles, seeing a "red article" and wanting to complete it... – p13

> The methods I used in 2015 are largely the same ones I use now, with the same mix of intentionally seeking out gaps, finding them in a more accidental way (like through a news item etc), or working through communally created lists like [Women in Red]. – p18

Using a Wikiproject's red link list as a starting point for content creation not only allows these editors to bypass the potentially lengthy process of identifying a missing article, but it also allows the Wikiproject–not the individual editor–to prioritize which content receives attention. In other words, a specific editor may begin an editing session by navigating to a Wikiproject's red link list without much thought as to how a particular page fits into the broader superset of missing content. Yet the links' presence on a Wikiproject red link list is due to a top-down, systematic effort to create content about a specific topic.

Figure 3.4. The first three sub-categories Wikiproject Women in Red's red link index. Red link lists are generated by hand (i.e. Crowd Sourced or CS) or generated programmatically by bots (i.e. Wikidata or WD).

While many personal to-do lists and Wikiproject red link lists are created "by hand", the Wikipedia community has created automated scripts–or bots–for programmatically generating and maintaining completeness lists. Editors primarily mentioned a bot named *Listeria*[28] throughout the study, which Wikiprojects such as Women in Red use to programmatically create some red link lists from Wikidata items. Listeria relies on existing structured Wikidata items to automatically create lists that conform to certain parameters, for instance an item's topical category or a person's gender. One of the bot's settings queries Wikidata for items that do not have corresponding articles in Wikipedia–these may be items that have been automatically added from external completeness lists or items that have articles in other languages–which produces list of red links that correspond to a certain topical category.

Creating red link lists programmatically and by hand both have unique advantages and disadvantages. Programmatic lists require less work to maintain and are easier to subset

[28] `http://magnusmanske.de/wordpress/?p=301`

by a variety of data properties. Furthermore, bots ensure completeness to the extent that the underlying dataset is complete; a script eliminates the possibility of omission or duplication due to human error.

However, at least in the case of Listeria, these programmatically generated lists are only as useful as the underlying dataset. Similar to a handcrafted list, editors must add Wikidata items for each subject, a process which may be less intuitive and accessible than simply creating a red link. Furthermore, structured data does not handle edge cases well–for instance, several editors mentioned the misgendering of non-binary and trans biographies–and incorrect entries are automatically propagated into Wikipedia without editor review.

3.7. Discussion

Developing a rich understanding of the editor community represents a critical first step towards any successful intervention on Wikipedia. While the Wikimedia Foundation can implement top-down solutions in order to reach its equity goals through tool development or social initiatives, it is ultimately the editors who create, improve, and maintain Wikipedia's content. Highlighting the editor community's key role on the platform and emphasizing the need for buy-in may seem self-evident, yet participant responses from this study illustrate disconnect between past Wikimedia efforts and the community they attempt to support.

> I would like to see less attempt to rule over day-to-day behavior on the site and more effort to serve as a support structure for the volunteers who are the glue that holds the project together. Too many junkets,

> too many engineers engaged in make-work projects of dubious value, not enough boots on the ground doing actual support work of the active volunteer community – p15

> At times, the WMF crowd had even expressed an attitude of contempt for "the community" usually in conjunction with getting pushback for certain WMF initiatives. I find it frustrating because the WMF crowd for the most part doesn't get into the actual work of editing (the time they had the bright idea that a computer could automatically generate articles...) and they don't understand the dynamics. – p10

While the above quotes do not necessarily represent the attitudes and opinions of the entire community–or even all editors in this study–they do reinforce prior instances of failed Wikimedia Foundation initiatives. The AFT's ultimate abandonment represents the most relevant example to this research, which we discussed in Chapter 2.

We conducted this study to better understand how editors currently work to fill knowledge gaps, and how interventions can support the community's efforts towards creating a more equitable Wikipedia. Broadly, our results confirm that, despite some disconnect between editors and the Wikimedia Foundation, these two entities share similar goals surrounding knowledge equity and representation; many of the editors in this study work primarily to create content with and about underrepresented populations, which is exactly the type of work the Wikimedia Foundation aims to make more effective. Yet in order to further the parallel goals of editors and of the Wikimedia Foundation, top-down interventions must be designed with the input of the editor community.

Our work advances prior research on knowledge gaps in several key ways. While prior studies have extensively quantified high profile content gaps (e.g. [**Redi et al., 2021, Wagner et al., 2015, Greenstein and Zhu, 2012, Hecht and Gergle, 2010a**]), few studies engage with the individuals and communities working to identify gaps and create content. Those studies that do examine editor behaviors tend focus on other facets of their experience, for instance examining editors' roles [**Welser et al., 2011**], how editors coordinate [**Viegas, 2007**], or the process by which readers become contributors [**Shaw and Hargittai, 2018, Preece and Shneiderman, 2009**]. To our knowledge, no prior work aims specifically to understand how editors conceptualize knowledge gaps and the workflows employed to both identify and create missing content.

Furthermore, our work reinforces a large body of empirical work examining potential causes of knowledge gaps. Editors reiterated that Wikipedia's policies tend to systematically exclude certain types of content (e.g. [**Gallert and Van der Velden, 2013, Wagner et al., 2016, Schneider et al., 2012**]), and that the contributor gap plays a major role in determining which types of content get added and improved [**Redi et al., 2021**]. While these studies use edit logs and other forms of trace data to analyse causes of content gaps, our research engages with the individuals who encounter these barriers in their work.

3.7.1. What is a Knowledge Gap?

RQ 1 aimed to better define Knowledge Gaps from Wikipedia's editors' points of view. Broadly, we found that editors frame knowledge gaps in terms of Wikipedia's scope: which subjects should be included, and how those subjects should be covered. A knowledge gap

could be missing content, or content that should exist but does not. A knowledge gap could also be the converse–point of view content, biased content, or content that does exist but should not.

By comparison, the second draft[29] of A taxonomy of knowledge gaps for wikimedia projects produced by the Wikimedia Foundation's research team defines knowledge gaps as "...disparities in content coverage or participation of a specific group of readers or contributors." [**Redi et al., 2021**] Later in the document the authors clarify their definition, focusing on "gaps".

> *A gap corresponds to an individual aspect of the Wikimedia ecosystem—for example readers' gender, or images in content—for which we found evidence of a lack of diversity, or imbalanced coverage across its inner categories (for example, proportion of readers who identify as men, women or non-binary in the case of the reader gender gap)*

The right side of Figure 3.5, reproduced from the taxonomy, visually illustrates this definition. Content gaps make up one third of the of the taxonomy, but this definition also includes contributor gaps and reader gaps. Returning to the definition above, because there is a disparity between the numbers of readers who identify as men, women, or non-binary, the Wikimedia foundation defines this disparity as a knowledge gap, or more specifically the "reader gender gap".

Wikipedia's editors and administrators define knowledge gaps differently, both in terms of measurement and scope. Figure 3.5 visually illustrates the difference in scopes, where the right side shows the Wikimedia Foundation's definition, and the left side shows

[29]The second draft of this document was the most up-to-date at the time of writing. Subsequent drafts may update this definition.

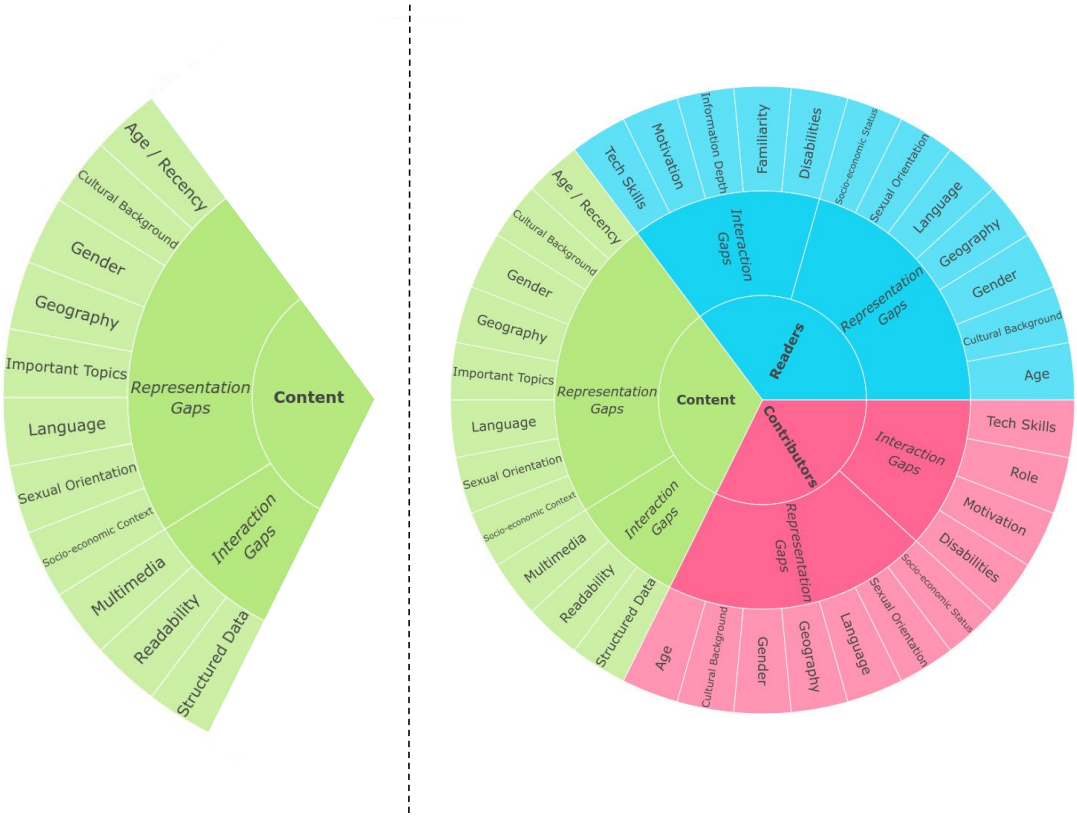

Figure 3.5. Editors conception of knowledge gaps (left) vs The Wikimedia Foundation taxonomy of knowledge gaps (right). Editors extensively discuss the contributor gap (the red portion of the figure), but frame the contributor gap as a *cause* of knowledge gaps rather than a knowledge gap itself. Reproduced from `https://meta.wikimedia.org/wiki/Research:Knowledge_Gaps_Index/Taxonomy`.

the subset of gaps that fit editors' definitions reported in this study. Editors acknowledge the existence of the contributor gap and its role in creating inequity, yet at no point during the study did they directly define the contributor gap as a knowledge gap.[30]

[30]In the author's opinion, categorizing contributor gaps and reader gaps broadly as "knowledge gaps" is a misnomer that leads to confusion. The Taxonomy of Knowledge Gaps is actually a Taxonomy of Wikipedia's Equity Problems.

Furthermore, the Wikimedia Foundation defines knowledge gaps in terms of equity, diversity, and balance. Gaps are measured in terms of internal disparities between categories. To return to the prior example, the difference in numbers of biographies about men, women, and non-binary people indicates the existence of a gender gap. In contrast, editors define knowledge gaps through an overarching concept of scope, bound–albeit imprecisely– by WP:Notability, WP:Reliability, and WP:Verifiability. Gaps are measured by their relationship to the superset of notable subjects that should exist within Wikipedia.

Both definitions have unique advantages and drawbacks. The Wikimedia Foundation's definition makes gaps relatively easy to quantify as long as researchers can define a category (e.g. gender) and a metric (e.g. number of articles), but makes the assumption that we should aim for equality across all categories and metrics. The editor's definition avoids the assumption of equality but makes measuring completeness impossible or at least intractable for the large number of categories where the superset of "all notable subjects" is imprecisely defined.

The goal of this study is not to pass judgement about which definition should be adopted, but we argue that it would benefit both editors and administrators to reach some consensus. As one editor remarked, even the term "knowledge gap" may be associated with the top-down initiatives that raise skepticism among many editors.

> "Knowledge gap" is lingo that comes from some of the WMF-based initiatives. I don't know if it comes in from academe or what, but I've seen the phrase used for the last 5-6 years. – p10

3.7.2. Providing Support for "Boots on the Ground"

RQ 2 and RQ 3 explored how editors identify and triage knowledge gaps. RQ 4 then imagined how social or technological interventions could make these processes more effective. In other words, how do editors decide where to allocate their efforts when producing content, and how can technology make this easier? In retrospect, RQs 2, 3, and 4 assume that knowledge gap identification is the preeminent barrier facing editors and the larger socio-technical system that is Wikipedia, at least with respect to creating equitable knowledge representation. This assumption makes some sense given the Wikimedia Foundation's published work framing the problem of knowledge equity; *identifying knowledge gaps* and *measuring knowledge gaps* are steps 1 and 2 the Foundation's 2030 strategic direction published in 2019 [**Zia et al., 2019c**].

Given our overarching goal is to better understand editor behavior, we would be remiss to focus entirely on knowledge gap identification when many participants highlighted other barriers and potential solutions in their editing workflows. For many editors, there exists more than enough missing content to create, the hard part is creating that content.

> OK, so let's suppose we have compiled and published some giant list of knowledge gaps within [the] scope of Wikipedia. Then what? Well, I'll look at all those football statistics that are missing and go "so what?" and keep on with my normal editing of my state's history and geography. At the end of the day, a volunteer-written encyclopedia reflects the interests of those who write it. – p4

Returning to Figure 3.1 and Figure 3.2–our simplified editing workflows–many editors suggested improvements for *gathering sources* and *writing articles*. For instance, editors

suggested that the Wikimedia Foundation provide easier access to paywalled or hard to find sources, and work with experts to update the WP:Notability, WP:Reliability, and WP:Verifiability policies that effectively maintain biases against underrepresented groups. Efforts such as the Wikipedia Library[31] and the Universal Code of Conduct [**noa, 2021**] represent initial steps towards these suggestions, which aim to address knowledge gaps created by Wikipedia policy.

Participants also suggested increasing support for both power users and newcomers. Suggestions for power users included paid support for editing, an expansion of the Wikimedian-in-Residence program[32], and various forms of an expert clearing house, where editors could request access to sources, collaboration on content creation, or help with technical questions. Though editors also acknowledged the value of newcomer support and outreach, these suggestions were less concrete, perhaps due to the high experience level of participants in this study. Editors recognized the importance of outreach programs like edit-a-thons and collaborations with universities, but also lamented the lack of newcomer retention and failure to convert newcomers to experienced, productive editors. Taken together, these suggestions aim to increase volunteer capacity, which would address knowledge gaps created by the contributor gap.

3.7.3. Knowledge Gap Identification

Despite the variety of proposed solutions that do not directly aim to identify missing content, we argue that systems for knowledge gap identification and quantification have a role to play in creating equitable knowledge representation. Wikipedia editors may

[31] `https://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Library`
[32] `https://en.wikipedia.org/wiki/Wikipedian_in_residence`

have more than enough content to produce and not enough time or resources to produce all of that content, but responses from participants in this study show that editors are willing to focus on certain topics if those topics are deemed sufficiently important. The success of projects such as Women in Red indicate that, while editors generally work on what interests them, systematized efforts to address underrepresented topics can direct the attention of at least a subset of editors to create tangible improvement in that topic's representation.

The key to successful systematic efforts that produce equitable information representation is creating interventions that fit into Wikipedia's existing social and technical framework. From a technical standpoint, we know that editors use markers (e.g. red links, stubs, and templates) to indicate areas of the encyclopedia that need improvement, and they use aggregated and curated lists of those markers (e.g. red link lists) to systematically create and improve content about specific topics. A hypothetical system aimed at identifying and mitigating knowledge gaps should use existing markers, such as red links or templates, to indicate subjects that could use attention. Listeria provides one example of successful integration which could be emulated by future projects; the bot programmatically produces red link lists of the same form editors and wikiprojects use for to-do lists. Throughout this study, editors suggested that more nuanced markers and more powerful systems that aggregate those markers could improve the missing article identification step in their workflows.

Integration into existing social frameworks represents a second key component of successful systematic content creation efforts. While tools such as Listeria provide powerful methods for aggregating markers, discovering and using these tools can be challenging,

even for advanced editors. Wikiprojects provide a centralized organizational structure that directs the efforts of groups of editors, and tools built to identify and fill knowledge gaps should leverage this structure. Wikiprojects both indicate that a topic is sufficiently important to deserve organized attention, and they can attract editors who are interested in creating content about that topic.

Women in Red is an example of successful integration between Wikipedia's social structures and technical affordances in order to identify, quantify, and mitigate a knowledge gap. The wikiproject uses Listeria to identify missing content, but creates familiar and easy to use red link lists to indicate where editors should direct their attention. As such, the majority of editors can create content with no investment into learning tools or maintaining personal curated completion lists, but can nevertheless participate in a systematic effort to create a more equitable version of Wikipedia. Future tools for identifying knowledge gaps could use Women in Red's use of Listeria as a model for successful integration between Wikipedia's social framework and technical affordances.

3.7.4. Reader-Sourcing

Among the variety of solutions proposed by participants in this study, several editors mentioned some form of reader-sourcing as a method for identifying missing content. Notably, participants brought up reader-sourcing independently, without priming from the research team; our study prompts did not suggest interaction with the the broader reader community in any way.

> To truly answer the question about the knowledge gaps on Wikipedia, we probably should ask the readers to tell us "what information did you

> come here to find but didn't?". Then we sort them into the knowledge gaps that we think Wikipedia should address and the knowledge gaps that we don't think it should address ([e.g.] "What time does the next bus leave from my house bound for the airport?"). - p4
>
> Pie in the sky idea: What if google (other search engines) selective released their metrics of poor search results. That is where someone has searched for some information and sparse detail is returned...Would this info not be a good pointer for where Wikipedia needs to be expanded? - p12

As discussed in Chapter 2, the implementation details of reader-sourced knowledge gap identification are critical to the system's success. While participants in this study may theoretically support reader-sourcing, prior push back against the AFT provides a contradictory example, illustrating that many editors feel the extra labor cost incurred outweighs the potential benefit. In the following chapter (Chapter 4) we use data collected by the AFT to explore the feasibility of reader-sourcing, specifically by reducing the workload associated with using such a system.

CHAPTER 4

Study 2: Classifying Reader-Sourced Feedback for Knowledge Gap Identification

Reader-sourcing provides one possible solution to latent knowledge gap identification. In a reader-sourced system, Wikipedia's consumers submit suggestions that identify missing or biased content. Wikipedia's readers represent a far more diverse population than its editors, so these suggestions would theoretically highlight underrepresented topics where editors' self-focus bias has resulted in latent content gaps. Notably, the content gaps we aim to identify are only latent to the relatively homogeneous editor population, so leveraging Wikipedia's more diverse reader population would theoretically call attention to this subset of missing content.

Researchers at the Wikimedia Foundation have previously built and deployed reader-sourced systems, though not with the express purpose of identifying all knowledge gaps. As discussed in Chapter 2, the AFT solicited reader feedback in order to identify one specific type of knowledge gap–low quality content[1]–and to aggregate suggestions for article improvement [**Halfaker et al., 2013**]. However, while the AFT gathered helpful feedback from some users, the tool also collected a large quantity of vandalism and spam responses. Of the responses reviewed and marked by editors, roughly half were marked

[1]Although low quality content arguably represents one form of a content gap, the AFT was never framed as a knowledge gap identification system.

"unhelpful". In order to use reader feedback to improve articles, editors were therefore required to manually triage and filter a high volume of feedback.

Figure 4.1 illustrates this process visually, depicting a simplified workflow that editors used to select, triage, and ultimately improve articles using reader feedback submitted to the AFT. In this workflow, editors allocated time to reading and flagging unhelpful feedback in order to find feedback which could lead to article improvement. Given that editors have a finite amount of time to allocate towards Wikipedia, any time allocated towards reading and flagging unhelpful responses represents time taken away from improving articles.

A hypothetical reader-sourced system designed explicitly for identifying *latent* knowledge gaps would face many of the same obstacles as the AFT. Any successful reader-sourced system must reduce workload overhead resultant from low quality feedback. In this research we therefore classified reader responses from the AFT in order to surface useful feedback while minimizing any additional workload associated with triaging those responses. We choose to classify feedback generated by the AFT rather than producing our own dataset due to the high cost and relatively small benefit of developing a novel feedback collection system. Many of the helpful suggestions in the AFT identify knowledge gaps due to missing or biased content in low quality articles.

We develop a classifier to predict whether reader feedback will be helpful or unhelpful. Ultimately, we aim to reduce workload associated with triaging and flagging responses. Figure 4.2 depicts a hypothetical workflow in which we automatically classify and filter reader responses as helpful or unhelpful, thereby eliminating the need to manually flag responses. Note that this classification pipeline does not handle all workload imposed on

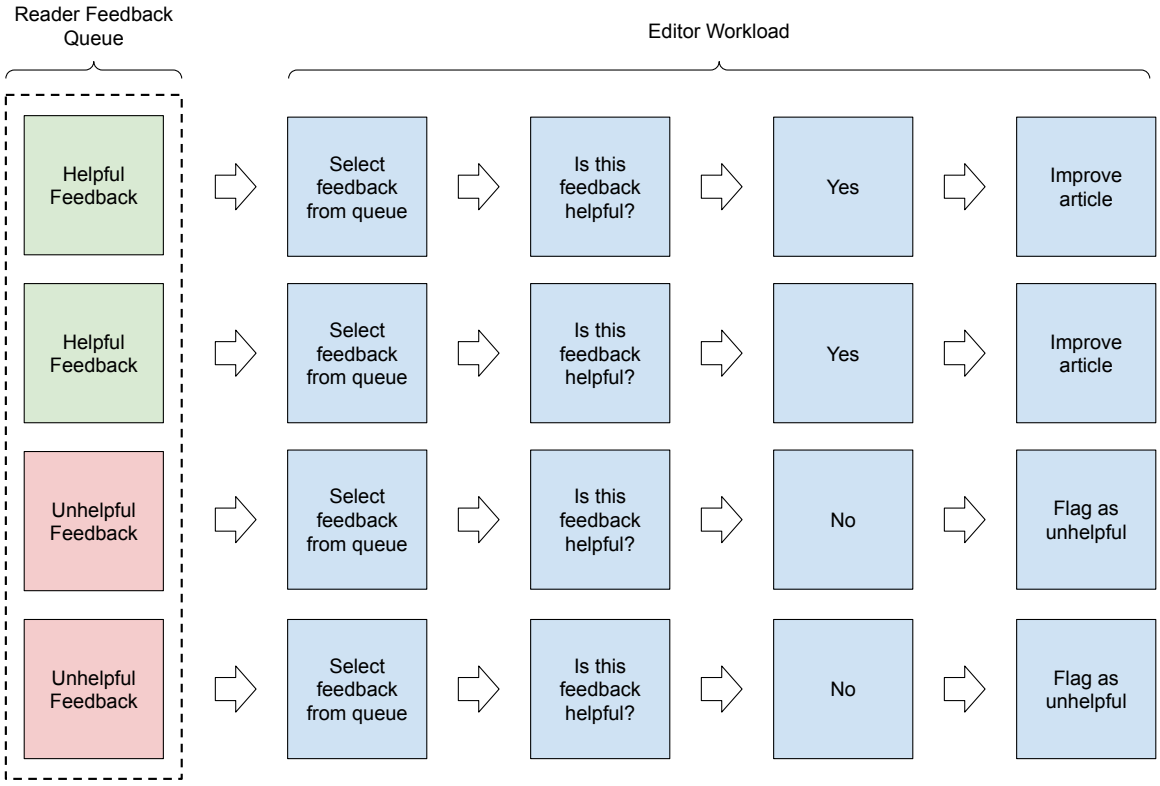

Figure 4.1. A simplified workflow that editors might use to select, triage, and ultimately improve articles using reader feedback submitted to the AFT.

editors by a reader-sourced tool; editors must read feedback, decide how best to respond, and ultimately allocate time towards improving the article. Nevertheless, the classifier does eliminate the most taxing and least productive component of the workflow, which ultimately led to the AFT's retirement.

Note also that we discuss workload at the editor population level, not the added workload for an individual editor. This distinction is important because we can not predict whether an individual editor will receive helpful or unhelpful feedback from the feedback queue. One editor might receive helpful feedback on his or her first try, while

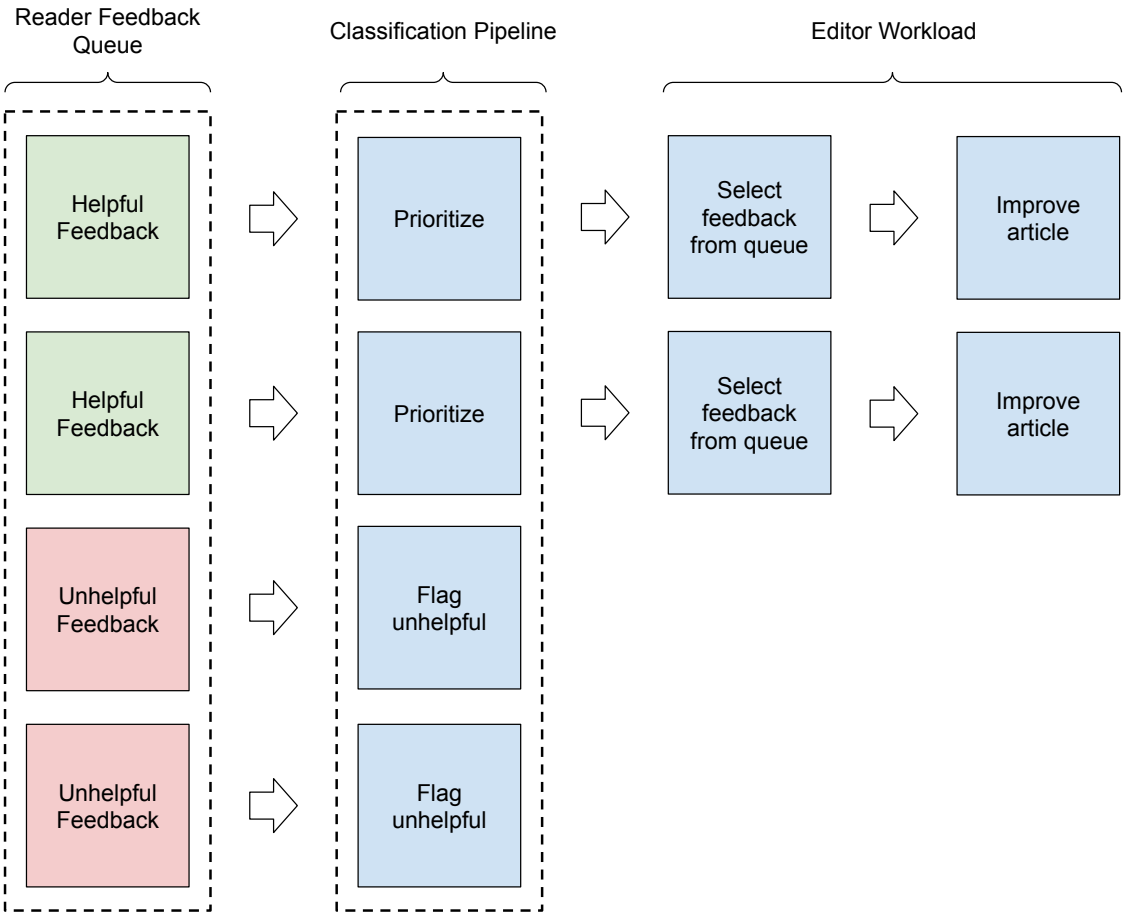

Figure 4.2. A hypothetical workflow in which we automatically classify and filter reader responses as helpful or unhelpful, thereby eliminating the need to manually flag responses.

a second less lucky editor might review 10 unhelpful feedback items before receiving one that is helpful. In this example, the Editor 1 would allocate no extra work to triaging and flagging unhelpful feedback, while Editor 2 would have to repeat the triage process ten times.

If we define workload at the population level, we assume that all reader feedback will have to be read and triaged by an editor at some point. While a classification system might reduce individual editor workload differentially–i.e. Editor 1 will receive no benefit while Editor 2 will benefit a great deal–on average the entire community will benefit from not triaging feedback.

We capture our goals of reducing editor workload to improve the feasibility of reader-sourced systems in the following research questions:

> ***RQ 1:*** *Can we build a machine classifier to predict whether reader feedback will be helpful or unhelpful to editors?*
>
> ***RQ 2:*** *To what extent can our machine classifier reduce overall workload associated with processing reader feedback?*

Our analysis took the form of a classification study using existing data and existing annotations. This portion of the study included 1) data processing, 2) classification, and 3) validation components. We then performed additional threshold and error analyses, which show that we can maintain high precision and still reduce the amount of feedback editors must triage at a rate of up to 66%. We describe each of these analyses in more detail below.

4.0.1. Dataset

The Wikimedia Foundation invested a significant amount of time and resources into developing and improving the AFT. The AFT was active for four years, and researchers developed and tested five different versions of the system. Given the investment into this system, we argue that developing and deploying a new data collection pipeline with the

Figure 4.3. The AFT's editor facing interface, showing feedback for the article "Golden-crowned Sparrow. Editors could mark reader feedback as *helpful* or *unhelpful*.

ability to collect data from authentic Wikipedia readers would provide at best a marginal improvement in data quality. Given that low quality content is a type of knowledge gap, findings from this study based on AFT data should be relatively transferable.

The original AFT dataset consists of 1,188,265 English comments (the full dataset is 1,549,842 comments, but also includes 298,749 French and 62,828 German comments). Data was collected between March 2013 and March 2014 during Version 5 of the AFT's deployment. The dataset is stored in a single CSV file, which is hosted on datahub.[2] Each comment contains the comment text and a binary field indicating whether the reader found what they were looking for. Each observation also includes metadata about the user who created the comment, the comment's creation data (e.g., timestamp) and associated article, and feedback that was submitted through the editor-facing AFT interface. Wikimedia provides the full schema[3] and interface design documentation.[4]

[2] `https://old.datahub.io/dataset/wikipedia-article-feedback-corpus`
[3] `https://www.mediawiki.org/wiki/Article_feedback/Version_5/Technical_Design_Schema`
[4] `https://www.mediawiki.org/wiki/Article_feedback/Version_5/Feature_Requirements`

X	mean	std	min	max
unhelpful	1.23	0.67	1	25
helpful	1.36	1.25	1	97

Table 4.1. Summary stats for all helpful and unhelpful variables, where helpful is limited to observations that received at least one helpful flag, and unhelpful is limited to observations that received at least one unhelpful flag. The majority of observations received only a single flag.

X	count	mean	std	min	max
net zero	9225	2.25	0.88	2	16

Table 4.2. Net Zero describes observations where helpful and unhelpful were both non-zero and summed to 0. Min and Max columns are the minimum and maximum number of flags a net zero observation received in the dataset. For example, the net zero observation with the maximum number of flags received 8 helpful and 8 unhelpful flags.

Included in the fields mentioned above, each observation in the AFT dataset contains metadata left by editors through the AFT's editor-facing interface. Most importantly for this study, editors could flag any reader feedback as *helpful* or *unhelpful* (as illustrated in Figure 4.3, though not all feedback was annotated. Editors were not required to flag feedback in order to use the AFT system, so unannotated feedback could be neutral (i.e. neither helpful nor unhelpful), or unannotated feedback could indicate that the feedback was never read. We discarded all unannotated feedback from our final dataset due to this ambiguity.

The AFT system aggregated and stored counts of helpful and unhelpful flags for each feedback item submitted by readers. For each observation we aggregate the helpful and unhelpful fields into a $netHelpful$ binary variable. For instance, one Wikipedia reader left the following comment about the Wikipedia page "Schizophrenia".

> [The article] gave me the base understanding of what schizophrenia is, though it would be nice to have had a "living with the illness" section.

11 editors flagged this comment as helpful and 1 editor flagged it as unhelpful, leaving a $netHelpful$ score of 1.

Because we cannot infer whether a zero indicates that the comment was neither helpful nor unhelpful or whether the comment was simply never read, we remove all comments where the sum of helpful and unhelpful flags is equal to zero. We choose a binary dependent variable over linear dependent variable[5] because it is not clear from the metadata why a specific observation might receive a greater number of flags. An observation might receive more helpful flags because it is indeed more helpful, but it also might receive more helpful flags because more editors viewed that particular feedback item, or because the editors viewing that feedback item were more predisposed to flagging feedback.

In order to limit the complexity of the classification task, we limit our dataset to the 1,188,265 English comments and exclude comments from the German and French dataset. Our final dataset consists of 105,760 observations, where 60,850 were labeled unhelpful and 44,910 were labeled helpful. Because our final dataset is relatively balanced (57.54% / 42.4%) we do not perform any additional sampling.

4.1. Feature Development

Our feature set consists of both lexical features extracted from comment text and a single feature extracted from metadata.

[5]In other words, we chose to predict the sign of $netHelpful$ instead of the value

4.1.1. Lexical Features

We test both term frequency-inverse document frequency (TF-IDF) bag of words (BoW) and word embedding vectorization for extracting lexical features from comment text. For both methods, we perform standard preprocessing steps, including punctuation removal, lowercasing, and tokenizing. We do not remove stop words for either vectorizer.

Our BoW vectorizer counts the occurrence of each word for a given piece of feedback and performs TF-IDF normalization over the feature matrix. TF-IDF normalization increases the weight of words that more uniquely characterize a specific observation, and decreases the weight of words that appear in many observations. We limit the vocabulary size to 10,000 words and exclude any word that appears in more than 90% of our observations. We use scikit-learn's preprocessor and TF-IDF vectorizer.

We initially attempted to re-use the embedding layer from Wikimedia's ORES Draft-Topic prediction model,[6] but this implementation resulted in poor classifier performance. Our baseline Logistic Regression classifier achieved an ROC-AUC score of just .53. We therefore create a custom word embedding vectorizer using Gensim's implementation of word2vec (w2v). We train a skip-gram model with a 100 dimensional hidden layer and limit the vocab size to 10,000. For each word in a given observation, we extract the 100 dimensional embedding and calculate the mean vector for the entire observation.

We ran an initial test between BOW and word embedding vectorization methods and found that word embeddings resulted in slightly higher classifier performance. For instance, our baseline Logistic Regression classifier achieved an ROC-AUC score of .63 vs .68 with BoW and w2v, respectively. This makes intuitive sense, as prior work indicates

[6]ORES word embeddings: `https://analytics.wikimedia.org/datasets/archive/public-datasets/all/ores/topic/vectors/enwiki-20200501-learned_vectors.50_cell.10k.kv`

that the embedding representation captures the higher level semantic meaning of a word [**Mikolov et al., 2013b, Mikolov et al., 2013a**]. For the remainder of the study, we use word embedding vectorization to generate our lexical feature vector.

4.1.2. Hand-Crafted Features

We extract a binary rating variable from each observation's metadata to improve the performance of our classifier. Before submitting feedback, each reader answered yes or no to the question "Did you find what you were looking for?" Since positive and negative comments are fundamentally different types of comments, this variable indicates whether the comment is likely to be positive or negative.

In future work it may be possible to engineer additional features extracted from each comment to improve model performance. For instance, we could include article level features (e.g. the article topic or quality) or the lexical similarity between a given comment and its associated article.

4.2. Classification

For our classification task we use the features described above to predict our single $netHelpful$ binary variable. We test a number of different classification algorithms to determine which provides the best performance. We use a traditional grid search approach based on a modified version of the hyperparameter grid used to produce the ORES EditQuality model.[7] This parameter grid includes configurations for Scikit-learn's implementations of Gradient Boosting, Random Forest, and SVM classifiers. Additionally, we add a logistic regression classifier to use as a baseline.

[7] `https://github.com/wikimedia/editquality/blob/master/config/classifiers.params.yaml`

We choose to represent RQ 1 as a single classification task due to the data format produced by the AFT. RQ 1 could be split into 2 distinct tasks–i.e. RQ 1.a: can we classify helpful reader feedback, and RQ 1.b: can we classify unhelpful reader feedback. A 2 task representation makes some theoretical sense given that each task is distinct. In a complete system, removing unhelpful feedback is a filtering task whereas promoting helpful feedback is a ranking task. However, the format of our dataset treats helpful and unhelpful as a single, binary variable. In other words, feedback that is not unhelpful is by definition helpful and vice versa. Furthermore, while we can rank feedback according to our confidence that it will be helpful, our binary helpful/unhelpful variable does not capture how helpful a certain feedback item might be. We therefore train a single classifier to predict whether feedback is helpful or unhelpful and leave a 2 task implementation to future work. We further explore the implications of this decision in our discussion.

4.3. Training, Validation, and Error Analysis

We split our dataset into 80/20 train and test subsets resulting in 84,608 comments for training and 21,152 comments for testing. We create representative samples for each in order to maintain the same ratio of positive and negative observations, where in both training and testing subsets the negative to positive observation ratio is 57/43. We then implement a 5-fold cross validation strategy on our training dataset in order to select the best classifier configuration from our parameter grid. We determine the best classifier by averaging the ROC-AUC score across all 5 folds for each configuration and selecting the highest mean score.

We retrain the best configuration on our entire training subset (all 84,608 comments) and validate this classifier against the previously unseen test subset. We report ROC-AUC, as well as precision, recall, accuracy, and F1 scores.

4.3.1. Threshold Analysis

Although under ideal conditions our system would perfectly classify positive and negative reader feedback, RQ 2 requires only that we reduce editor workload. From a population perspective, reducing editor workload associated with triaging feedback does not require perfect recall or specificity; by filtering out any negative examples or highlighting any positive examples we remove some amount of triage work that the editor population would have to do by hand.

Indeed, it may be more problematic or costly to misclassify a large quantity of feedback than to simply not classify ambiguous feedback [**Elkan, 2001**]. A successful system could flag the most unhelpful feedback and the most helpful feedback, while allowing editors to sort through feedback about which we are less certain.

We illustrate this visually in Figure 4.2 and Figure 4.4. As stated above, Figure 4.2 shows a hypothetical workflow in which our classifier perfectly predicts helpful and unhelpful feedback, thereby eliminating the need for editors to flag feedback manually. However, due to the imperfect performance of our model, some feedback will inevitably be misclassified. Figure 4.4 shows a modified version of this workflow, where we only predict a certain class when our confidence exceeds a certain probability threshold. In this workflow, editors manually triage feedback items for which our classifier reports low confidence

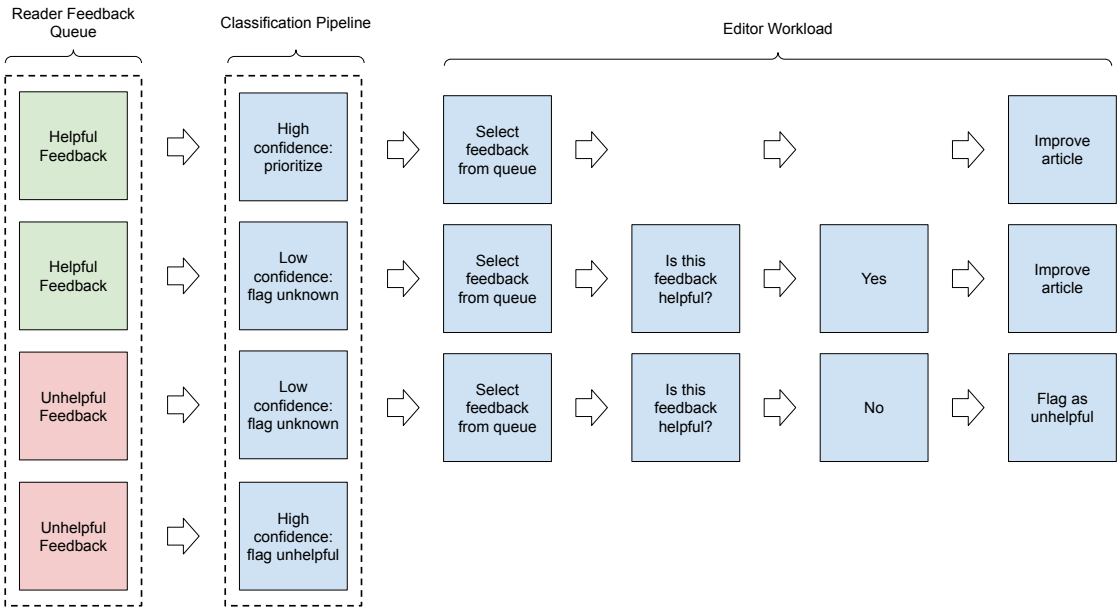

Figure 4.4. A hypothetical workflow in which we automatically classify and filter reader responses as helpful or unhelpful, but only for observations for which we can make a high confidence prediction.

(rows 2 and 3), but they do not have to triage feedback items for which we can make a confident prediction (rows 1 and 4).

In order to determine 1) how much we can reduce editor workload associated with triage and 2) what types of feedback we systematically classify either correctly or incorrectly, we perform a threshold analysis. At confidence levels from .5, .6, .7, .8, and .9 we classify any prediction for which we are less confident as "uncertain". For example, if for a given observation our classifier predicted a .75 probability of helpful, a .25 probability of unhelpful, and our threshold was .8, we would classify this observation as "uncertain".

For each threshold value we then report standard performance metrics. We also report several high level metrics, which we define as:

$$percentTriageReduction = \frac{truePositives + trueNegatives}{totalObservations} * 100 \tag{4.1}$$

$$percentMisclassified = \frac{falseNegatives + falsePositives}{totalObservations} * 100 \tag{4.2}$$

$$percentUncertain = \frac{unclassifiedObservations}{totalObservations} * 100 \tag{4.3}$$

Finally we perform an error analysis to determine if there are types of feedback we systematically misclassify at certain confidence thresholds. We use Empath's pre-built category model to categorize feedback. For each threshold value we create a subset of new misclassifications, or the subset of misclassifications that were classified as "uncertain" at the next highest threshold value. We then report the top lexical categories for each new misclassification subset at each threshold value. This error analysis allows us to determine whether we start to systematically misclassify certain types of feedback at lower confidence thresholds. For instance, if the empath category "feminine" appears when we decrease our probability threshold from .7 to .6, we might infer that at a .6 probability threshold we run the risk of misclassifying feedback about women (but we do not run the same risk at a probability threshold of .7).

4.4. Model Performance

Table 4.3 shows various performance metrics for the best configuration of each classifier. Best configurations were selected using the highest mean ROC-AUC score across 5-fold cross validation on a training set. Each model was then retrained on the entire

Classifier	Accuracy	Precision	NPV	Recall	F1	ROC-AUC
Gradient Boosting	0.659251	0.612059	0.688370	0.547898	0.578204	0.711357
Logistic Regression	0.642755	0.595956	0.669042	0.502844	0.545455	0.688940
Random Forest	0.656494	0.618645	0.676765	0.506189	0.556796	0.708075
SVC RBF	0.636528	0.602197	0.651753	0.434036	0.504472	0.682468
SVC Linear	0.634008	0.599404	0.649065	0.426341	0.498273	0.680173

Table 4.3. Classification performance metrics for the best hyper parameter configuration of each classifier. Best configurations were chosen using ROC-AUC values. NPV stands for Negative Predictive Value.

training subset and validated against held out test data. The table below shows each model's performance against the held out test dataset.

All 5 models performed similarly. Both SVM models performed the worst, with ROC-AUC values of .680 and .682. The Gradient Boosting and Random Forest classifiers performed the best, achieving slightly higher ROC-AUC scores of .711 and .708. Both SVM models had the lowest F1 values of .498 and .504, while the Gradient Boosting and Random Forest classifiers reported F1 values of .578 and .557.

Examining ROC, Precision-Recall, and Negative Predictive Value vs Specificity Curves Figures 4.54.6 illustrate tradeoffs between the rate of correctly classified examples and the rate of incorrectly classified examples. Each model provides the probability that a given example belongs to a given class. If we dictate that our classifier must exceed a certain probability threshold in order to make a prediction for a given example, both the number of correct and incorrect predictions will decrease.

Figure 4.5 compares ROC curves for all 5 models. Again, all 5 curves are relatively similar indicating that the predictive power of all 5 models differs very little. The Gradient Boosting and Random Forest classifiers both have a slightly higher true positive rate for a given false positive rate, which contributes to higher ROC-AUC scores. Similarly, while

both SVM classifiers perform similarly to the logistic regression model, at high false positive rates they both have a slightly lower true positive rate as illustrated by the small dip around .8. This characteristic explains both SVMs' lower ROC-AUC scores.

In the context of classifying reader feedback, comparing ROC curves indicates that the Gradient Boosting and Random Forest classifiers will perform better when flagging helpful reader feedback. This performance advantage will be more pronounced with lower false positive rates (and higher probability thresholds); if we require higher confidence from our classifier in order to make a prediction, the Gradient Boosting and Random Forest classifiers will correctly identify helpful feedback more frequently, and therefore flag unhelpful feedback as helpful less frequently. We can illustrate this visually by inspecting the area under each of the curves. As we travel from right to left along all 5 curves, the areas under the Gradient Boosting and Random Forest curves remain greater than the other 3. The Logistic Regression curve starts similarly to both the Gradient Boosting and Random Forest curves but falls off more quickly, resulting in a lower area at higher confidence thresholds. Both SVM configurations experience a small dip in performance around the .8 false positive rate, but then match the relative performance of the Logistic Regression. For the SVMs, both the dip and the consistently lower true positive rate compared to the Gradient Boosting and Random Forest classifiers indicate lower performance.

The relatively continuous nature of all 5 curves indicates the lack of an "obvious" confidence threshold value to choose. While an elbow or corner along the ROC curve would illustrate that lower threshold values lead to diminishing increases in the true positive rate, the relatively smooth shape of all 5 curves shows a relatively consistent relationship between true positives and false positives.

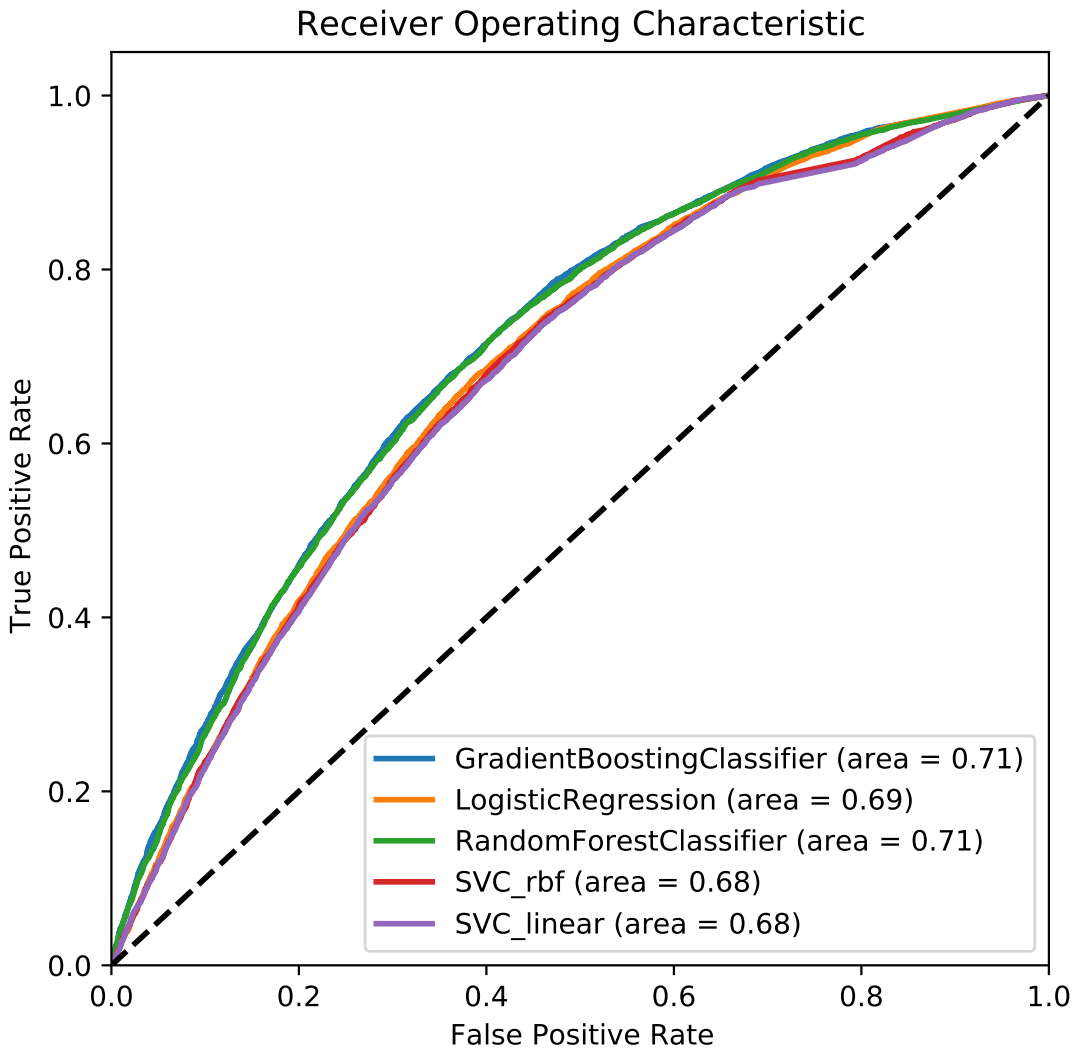

Figure 4.5. ROC curves and ROC-AUC values for the best hyper parameter configuration of each classifier.

A comparison of precision-recall curves in Figure 4.6 illustrates the same point. While the Gradient Boosting and Random Forest classifiers perform better than both SVMs and

the Logistic Regression for all threshold values (or lower recall values), the performance advantage increases at higher confidence thresholds.

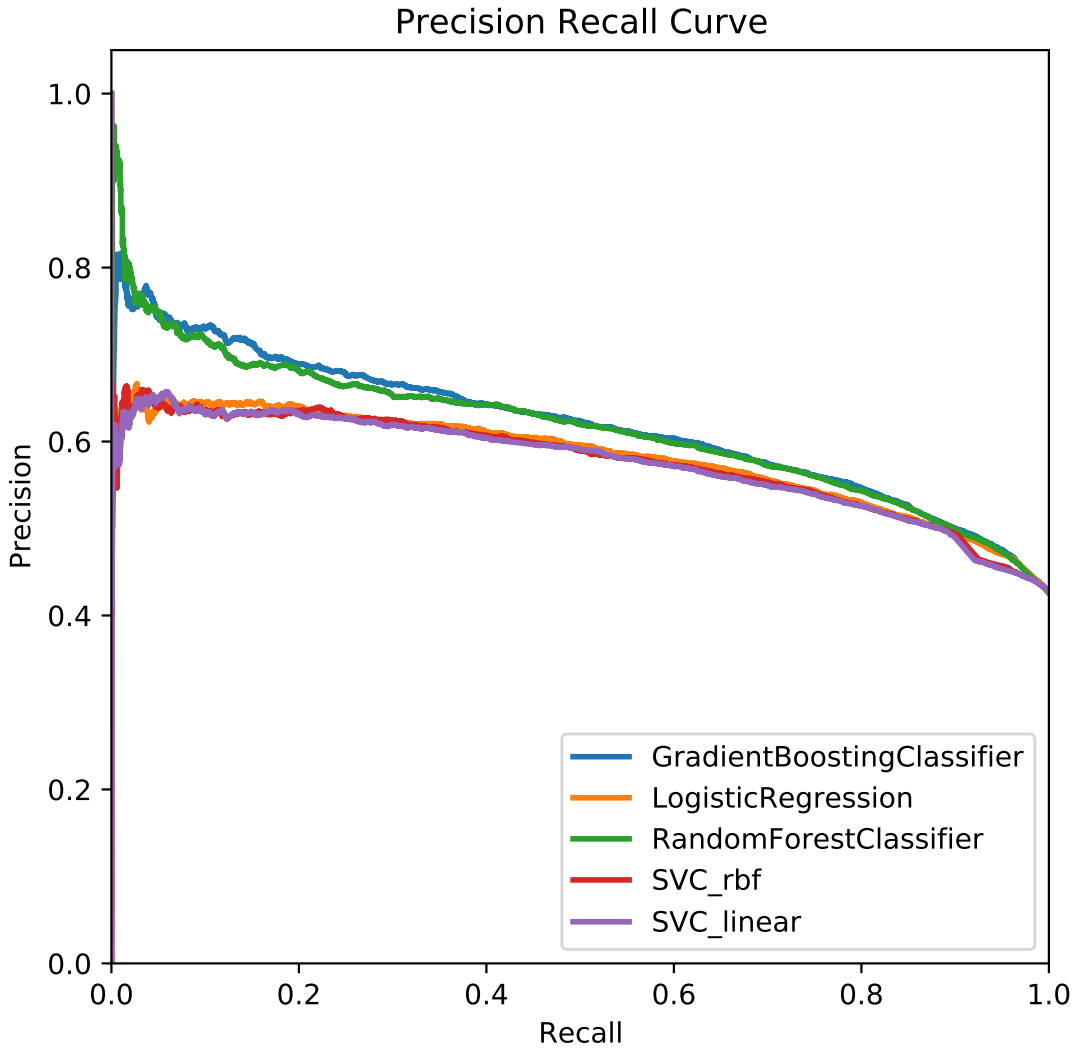

Figure 4.6. Precision Recall curves for each classifier.

The Negative Predictive Value vs Specificity[8] Curve (Figure 4.7) illustrates that the same relationship does not necessarily hold true for negative predictions. The Gradient Boosting, Random Forest, and Logistic Regression models all perform relatively similarly at all threshold values, while Both SVM classifiers show a slight dip in performance for high threshold values.

Taken together, the precision-recall and Negative Predictive Value vs Specificity Curves indicate that the Gradient Boosting and Random Forest classifiers primarily outperform other models when predicting positive examples with higher confidence thresholds. In the context of identifying helpful reader feedback, this means that all 5 models identify unhelpful feedback with similar accuracy. While requiring higher confidence in order to make a prediction does decrease the number of unhelpful comments flagged as helpful, classifier choice does not affect this relationship. Conversely, the Gradient Boosting and Random Forest classifiers both excel when identifying positive feedback when we require higher confidence to make a prediction. In this scenario they less frequently flag unhelpful feedback as helpful.

The following 5 graphs in Figure 4.8 compare precision-recall and Negative Predictive Value vs Specificity Curves for each classifier. These graphs show that, while the shape of each curve may differ slightly, all 5 models classify negative feedback more accurately than positive feedback for any given threshold value.

Comparison between ROC, Precision-Recall, and Negative Predictive Value vs Specificity Curves for all 5 classifiers indicates that the Gradient Boosting classifier outperforms the other 4 models for all of our chosen performance metrics. Other considerations such

[8]The Negative Predictive Value vs Specificity curve is the equivalent of the Precision-Recall curve for negative examples

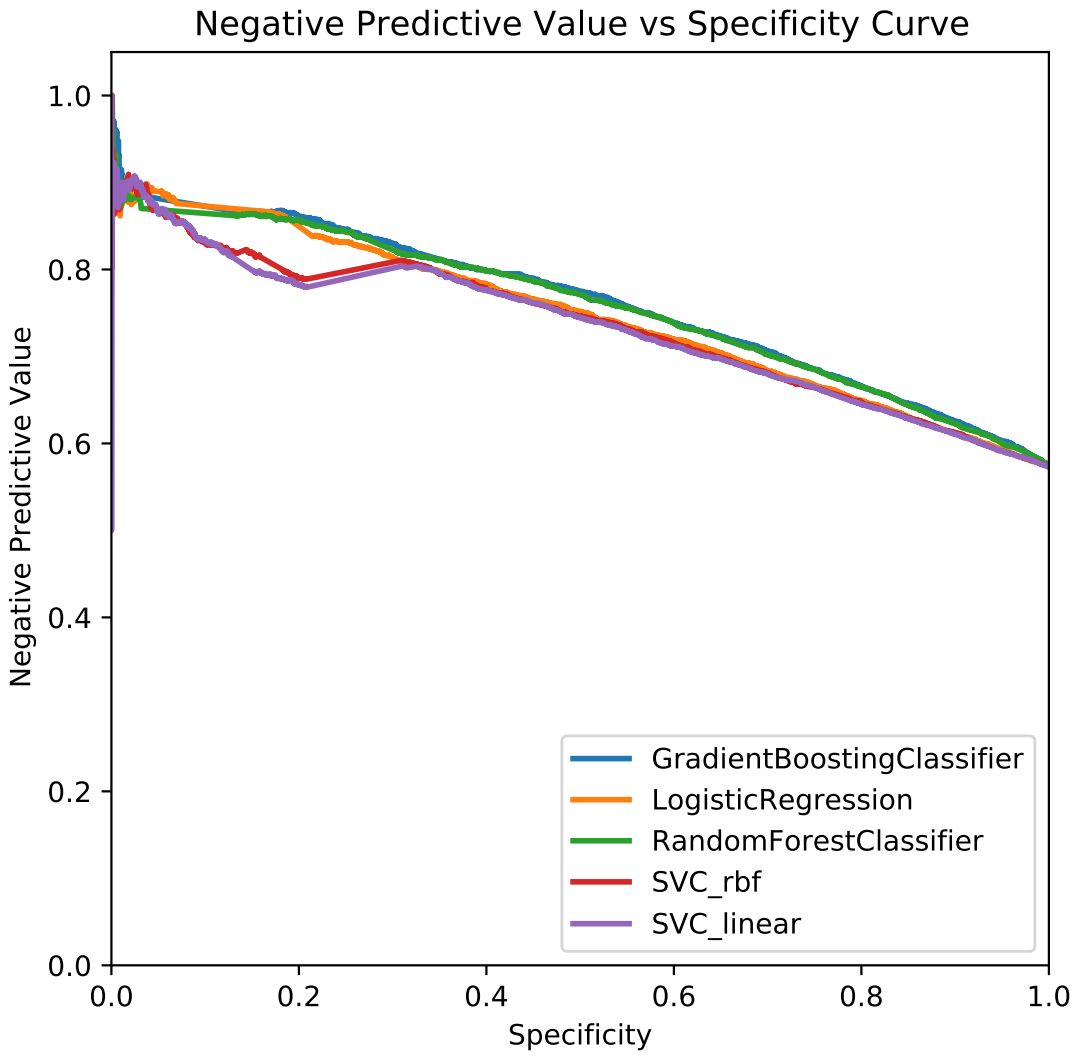

Figure 4.7. Negative Predictive Value vs Specificity curves for each classifier. The Negative Predictive Value vs Specificity curve is the equivalent of the Precision-Recall curve for negative examples.

as training time or prediction time could favor other models, but for the purposes of this analysis we focus on classification performance. Given the Gradient Boosting classifier's

ability to more accurately predict reader feedback under most conditions–and especially for higher precision and low recall scenarios–in the following discussion we only examine this model.

4.4.1. Threshold analysis

In the following analysis we explicitly focus on the thresholding approach discussed in the above comparison of ROC, precision-recall, and Negative Predictive Value vs Specificity Curves. While our best ROC-AUC value of .711 indicates relatively mediocre classification performance [**Metz, 1978**][9], RQ 2 dictates that we reduce editor workload, and not that we produce a perfect classifier. If we adjust the probability threshold that we require in order for our classifier to make a prediction we can still reduce editor workload (albeit by a smaller margin) while also reducing the number of positive and negative misclassifications. Returning to Figure 4.4 which depicts our hypothetical workflow, by increasing the probability threshold we increase the number of observations marked "unknown" (rows 2 and 3), but we reduce the number of misclassified observations.

The following table shows performance metrics for thresholds of .5 through .9 in .1 step increments. At a threshold of .5, the model performs as a normal binary classifier. All examples are assigned a prediction, where probabilities of above .5 are assigned to the positive class, and probabilities less than .5 are assigned to the negative class. At higher probability thresholds the classifier assigns examples with probabilities above the threshold or below 1 minus the threshold to the positive or negative class, and leaves examples that fall between these two values as unknown. For instance, at a threshold of

[9]ROC-AUC interpretations are highly context and domain dependent (as we show in the following two sections), but as a general heuristic many papers report .7 to .8 as "fair" performance.

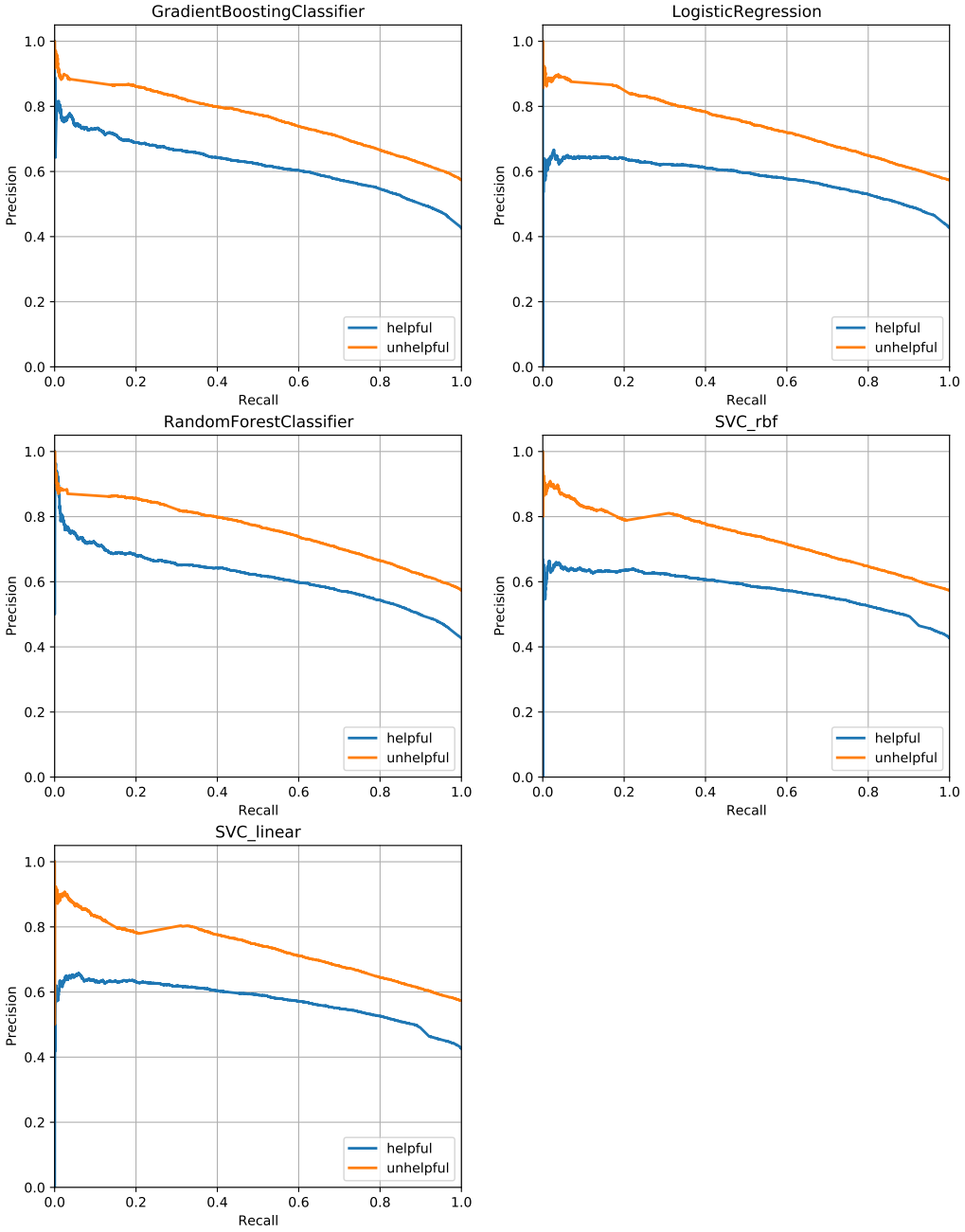

Figure 4.8. Comparison of Precision-Recall and Negative Predictive Value vs Specificity curves for each classifier.

.7, a probability of .8 would be predicted as positive, a probability of .2 would be predicted as negative, and a probability of .5 would be left unclassified.

Standard classification performance metrics (e.g., accuracy, precision, negative predictive value, recall, specificity, F1, and ROC-AUC) are all calculated after removing all unclassified examples from the test set. Additionally, we define several metrics that capture the performance of the classifier in the context of triaging reader feedback. We define *precentTriageReduction* (Equation 4.1) as the number of correct classifications over the total number of examples in the test set (including unclassified examples). In concrete terms this is the percent of classifications editors no longer need to triage because they have been removed from the pipeline or bubbled up as helpful. We define the *percentMisclassified* (Equation 4.2) as the number of false positives and false negatives over the total number of examples (including unclassified examples). *percentUncertain* (Equation 4.3) is the number of unclassified examples over the total number of examples, or the percent of feedback that would still need to be reviewed by an editor.

As we increase the probability threshold from .5 to .9, accuracy, precision, negative predictive value, and specitivity all increase, while recall, F1, and ROC-AUC decrease. Higher accuracy, precision, and negative predictive value all indicate that when the classifier makes a prediction, it tends to predict both positive and negative examples correctly more often. The increase in specificity combined with the decrease in recall, F1, and ROC-AUC indicates that as we increase the threshold value, the classifier increasingly predicts feedback as negative or leaves feedback as unclassified. For instance, at a threshold value of .8 the classifier only predicts 5 examples as positive (4 correctly and 1 incorrectly), and at a threshold value of .9 it does not predict any positive feedback.

Threshold	Accuracy	Precision	NPV	Recall	Specificity	F1	ROC-AUC
0.5	0.659251	0.612059	0.688370	0.547898	0.741984	0.578204	0.711357
0.6	0.730130	0.669869	0.755543	0.536096	0.844403	0.595563	0.746551
0.7	0.799775	0.746218	0.805447	0.288874	0.967707	0.416510	0.702323
0.8	0.858755	0.800000	0.858856	0.009615	0.999601	0.019002	0.541056
0.9	0.894737	0.000000	0.894737	0.000000	1.000000	0.000000	0.643428

Table 4.4. Standard performance metrics at confidence thresholds ranging from .9 to .5 for the Gradient Boosting classifier.

As we increase the probability threshold from .5 to .9, our percent triage reduction and percent misclassified metrics decrease, while our percent uncertainty increases. Logically this makes sense; when we require higher confidence to make a prediction the amount of feedback left unclassified increases, which means the amount of work required to triage unclassified feedback also increases. However, because our classifier makes predictions with higher accuracy, the number of misclassifications also decreases.

In the context of triaging reader feedback, as we increase our probability threshold we can still save editors a considerable amount of work while reducing the number of misclassifications. For instance, at a threshold of .7 we misclassify only 6 percent of reader feedback, but we reduce editors' workload by about 24 percent. If we reduce the threshold to .6 we misclassify 16 percent of reader feedback, but we reduce editor workload by 44 percent. Across our 21036 example test dataset, work reduction from a .7 or .6 probability threshold translates to 4969 or 9223 feedback items that editors no longer have to triage. If we extrapolate out to our entire train and test dataset, this work reduction would translate to 46114 or 24844 feedback items respectively that no longer need triaging.

Threshold	triageReduction	Misclassified	Uncertain
0.5	0.659251	0.340749	0.000000
0.6	0.438439	0.162056	0.399506
0.7	0.236214	0.059137	0.704649
0.8	0.119367	0.019633	0.861000
0.9	0.007273	0.000856	0.991871

Table 4.5. Additional metrics at confidence thresholds ranging from .9 to .5 for the Gradient Boosting classifier. At high confidence thresholds most observations are marked unknown.

4.4.2. Topic Analysis

In order to better understand classifier performance, we investigate the topical makeup of both correct and incorrect classifications. We report the top five comments in Table 4.6 and Table 4.7. Using a dictionary based approach, we generate the top 5 topics at each probability threshold for all four quadrants of the confusion matrix: true positives, true negatives, false positives, and false negatives. We use the Python library *Empath* to generate categories at each threshold level.

By comparing topics across these 4 quadrants we can determine whether misclassifications systematically affect certain topics, or whether these errors are relatively evenly distributed. Systematically misclassifying specific types of feedback could introduce more bias, which would ultimately negate the end goal of this project. Conversely, understanding what types of content are being correctly classified may help us understand where the classifier excels. Returning to our previous example, if the empath category "feminine" appears in the false negative column, we risk systematically misclassifying useful feedback about female subjects. If the category swear-words appears in the true negative column, we might infer that we can effectively filter abusive responses.

Overall, we observe relatively high overlap between topics across thresholds and quadrants of the confusion matrix. Of the 20 potential cells in the confusion-topic matrix (Table 4.6), 'internet' occurs 14 times, 'communication' occurs 13 times, 'writing' occurs 10 times, and 'reading' occurs 9 times. Notably, none of these top topics appear for negative predictions when our confidence threshold is set to .9, or positive predictions when our confidence threshold is set to .8 (recall that our classifier does not make any positive predictions with above .9 confidence). This likely indicates that these common topics are too general to help our classifier make a confident prediction.

Some topics appear consistently in either positive or negative predictions. Journalism appears at the .5 through .7 confidence thresholds for both true and false positive predictions, but in none of our negative predictions. Similarly, negative emotion appears at all thresholds for true negatives, at the .7 and .8 thresholds for false negatives, and in none of the positive predictions. These common topics may be strong predictors of either positive or negative feedback, but their appearance in both the true and false rows of the confusion matrix indicates the highly contextual nature of the concept of helpfulness when applied to reader feedback. For instance, while negative emotion appears in most unhelpful feedback, it also appears in some helpful feedback that our model misclassifies as unhelpful. Similarly, while the topic journalism appears in most helpful feedback, our model tends to misclassify unhelpful feedback that falls under the category journalism.

Notably, the predictions about which our classifier is most confident also have a higher number of unique topics. This could be because these topics are strong predictors of helpful or unhelpful feedback and therefore result in high confidence prediction, or simply it could result from a low number of predictions which produces a somewhat random

Threshold	TP	TN	FP	FN
0.5	internet	communication	internet	communication
	communication	internet	communication	internet
	writing	writing	writing	writing
	journalism	reading	journalism	reading
	reading	negative-emotion	reading	speaking
0.6	internet	internet	internet	internet
	communication	communication	communication	communication
	writing	negative-emotion	writing	writing
	journalism	writing	journalism	reading
	reading	messaging	reading	business
0.7	internet	negative-emotion	communication	internet
	communication	internet	internet	communication
	writing	communication	writing	negative-emotion
	journalism	messaging	reading	computer
	reading	phone	journalism	messaging
0.8	speaking	negative-emotion	help	internet
	help	internet	office	negative-emotion
	office	hate	dance	phone
	dance	communication	money	social-media
	money	phone	wedding	computer
0.9	None	negative-emotion	None	sexual
		domestic-work		domestic-work
		family		family
		home		swearing-terms
		shopping		ridicule

Table 4.6. Top terms for feedback classified at each confidence threshold for true positives, true negatives, false positives, and false negatives.

distribution of topics. For instance, 'domestic_work' and 'family' appear for both true negatives and false negatives, but only at the .9 confidence threshold. Family, home, and shopping appear only for true negatives at the .9 confidence threshold, and 'sexual', 'swearing_terms', and 'ridicule' appear only for false negatives at the .9 confidence threshold. 'Help', 'office', 'dance', and 'money' all appear only for true and false positives at the .8 confidence level.

We perform a second, similar analysis where we analyze the top 5 topics for the difference between confidence thresholds. For instance, the topics in the .8 - .9 row are topics that appear in feedback for which our predictions are only 80 to 90 percent confident. This analysis would indicate whether decreasing our confidence threshold systematically incorrectly classifies certain types of feedback. For instance, the decrease from .7 to .6 confidence level results in false negatives that fall under the 'business' category, even though 'business' does not show up in any other cells of the confusion-topic matrix. We might therefore extrapolate that decreasing the threshold from .7 to .6 biases our results against business related feedback.

Comparison between both confusion-topic matrices shows that decreasing the confidence threshold rarely results in biases against specific topics. For instance, 'friends' appears in the .8 to .7 false negative category and 'business' appears in the .7 to .6 false negative category, but otherwise topics in both tables match relatively well. This likely indicates that reducing the confidence threshold does not result in systematic bias towards or against certain types of feedback, but instead broadens the topics of feedback that are both correctly and incorrectly classified in a relatively distributed way.

4.5. Discussion

In this study we aimed to classify helpful and unhelpful reader-sourced feedback in order to reduce editor workload associated with feedback triage. We developed and tested several machine classifiers that use a combination of word embeddings and hand crafted features, ultimately achieving a maximum ROC-AUC value of .71. Furthermore, using a

Threshold	TP	TN	FP	FN
0.5-0.6	internet communication writing journalism reading	communication internet writing reading speaking	internet communication writing reading journalism	communication internet writing reading speaking
0.6-0.7	internet writing communication journalism reading	communication internet writing reading speaking	internet communication writing journalism reading	internet communication writing reading business
0.7-0.8	internet communication writing journalism reading	internet communication negative-emotion writing messaging	communication internet writing reading journalism	internet communication friends messaging negative-emotion
0.8-0.9	speaking help office dance money	negative-emotion internet communication phone hate	help office dance money wedding	internet negative-emotion phone social-media computer
0.9-1.0	None	negative-emotion domestic-work family home shopping	None	sexual domestic-work family swearing-terms ridicule

Table 4.7. Top terms for *additional* feedback classified at each lower confidence threshold for true positives, true negatives, false positives, and false negatives. This table differs from Table 4.6 in that each row's top terms are calculated only from the difference in threshold levels.

threshold approach we show that we can reduce the quantity of feedback editors review while maintaining relatively high precision, as demonstrated in Table 4.4 and Figure 4.10.

As noted above, the classifier's relatively unimpressive performance scores (e.g. ROC-AUC and F1) [**Metz, 1978**] obscures its real world benefits. The primary problem with the final iteration of the Article Feedback Tool was not that editors could not determine

whether feedback was helpful or unhelpful, but rather that the process of triaging the large volume of feedback created an overwhelming amount of additional work for already time constrained editors. Therefore, simply reducing the amount of feedback that requires triage could result in a feasible reader sourced system.

4.5.1. All (Mis)classifications are not Created Equal

In the context of triaging reader feedback, not all quadrants of the true/false/positive/negative confusion matrix carry the same real world implications. Both true positives and true negatives are feedback items that editors no longer have to triage and therefore result in work reduction. A true negative would be removed from the comment queue and would never be seen by the editor, requiring no additional action. In the case of a true positive, an editor would receive helpful and actionable feedback from the comment queue which might require research or editing work. These are both ideal scenarios for leveraging reader feedback; editors receive helpful suggestions for article improvements while never interacting with spam, vandalism, or otherwise unhelpful feedback.

A false positive results in a similar scenario as the original AFT platform. The unhelpful feedback would be misclassified as helpful, and the editor must therefore manually discard it from the comment queue, resulting in additional work for the editor.

False negatives represent a special case and are potentially the most problematic. In the case of a false negative, the model would misclassify helpful feedback as unhelpful and automatically discard the helpful feedback from the comment queue. Editors would therefore never see and triage these particular types of feedback. In the context of bias

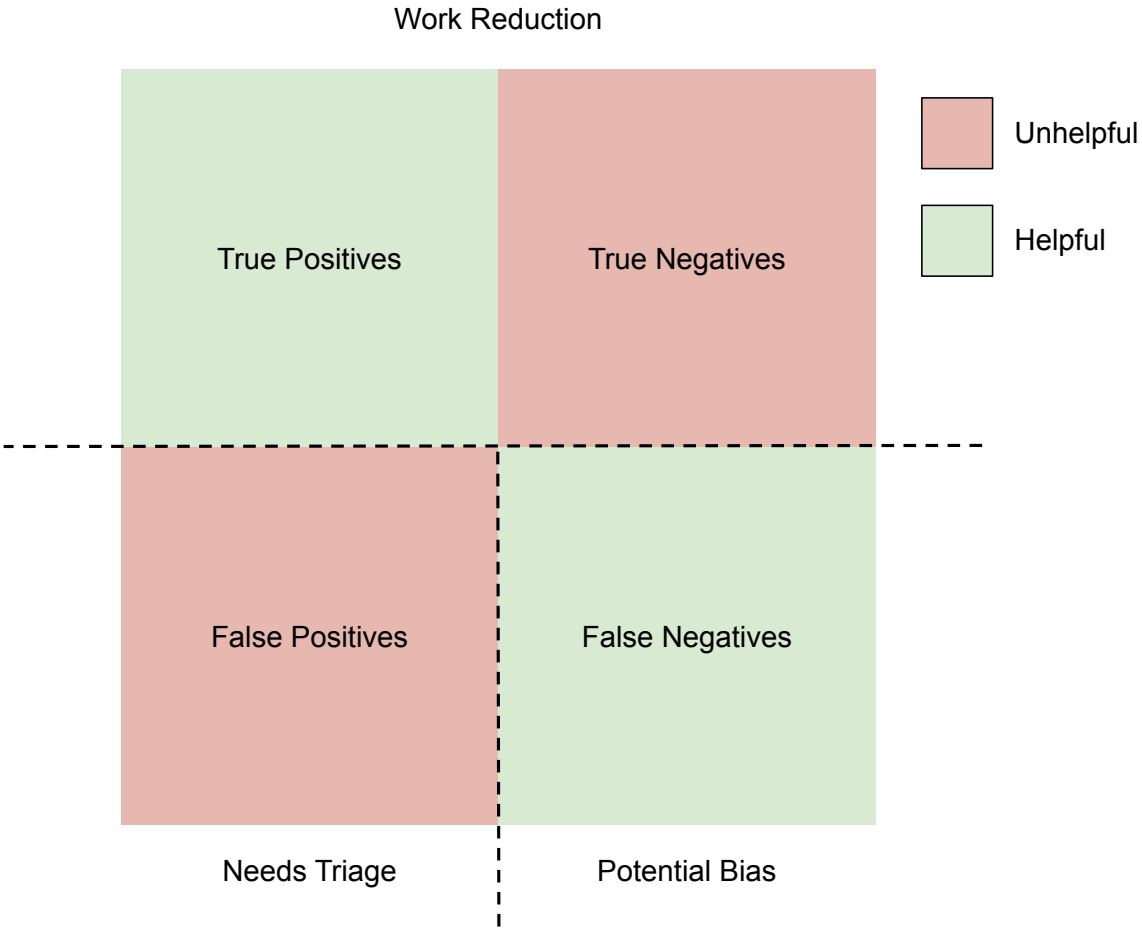

Figure 4.9. Implications for each cell of the confusion matrix. True Positives and True Negatives contribute to work reduction. False positives will still be included in the feedback queue and therefore must be triaged by editors. False Negatives represent useful feedback that will be hidden from editors by the system.

identification, this behavior becomes problematic if certain types of feedback are systematically misclassified as negative. These systematic misclassifications would introduce bias into the system by "hiding" certain topics rather than surfacing potential biases. We

Threshold	Unseen
0.5	0.192717
0.6	0.103252
0.7	0.051959
0.8	0.019585
0.9	0.000856

Table 4.8. The percentage of false negatives at certain thresholds. A false negative is helpful feedback that has been misclassified as unhelpful and effectively hidden from editors.

can examine the overall percentage of these potentially problematic false negatives with Equation 4.4, which we calculate for confidence thresholds of .5 through .9 in Table 4.8

$$percentUnseen = \frac{falseNegatives}{totalObservations} * 100 \tag{4.4}$$

Unfortunately, as we increase our confidence threshold to reduce our misclassification rate, our classifier tends to bias towards false negatives instead of false positives. Figure 4.10 illustrates this trend with converging *percentUnseen* and *percentMisclassified* lines. In other words, as we increase our confidence threshold, our errors tend to bias towards "hiding" feedback from editors that is actually helpful rather than including feedback that is unhelpful. As the probability threshold increases the overall number of errors decreases, but the proportion of potentially problematic errors from a systemic bias standpoint tends to be higher.

This pattern indicates that we need to be cognizant of the topical areas of feedback that we misclassify. Fortunately, our high level topic analysis indicates that false negatives at high confidence thresholds are predominantly characterized by the topics sexual, domestic work, family, swearing terms, and ridicule, none of which immediately suggest

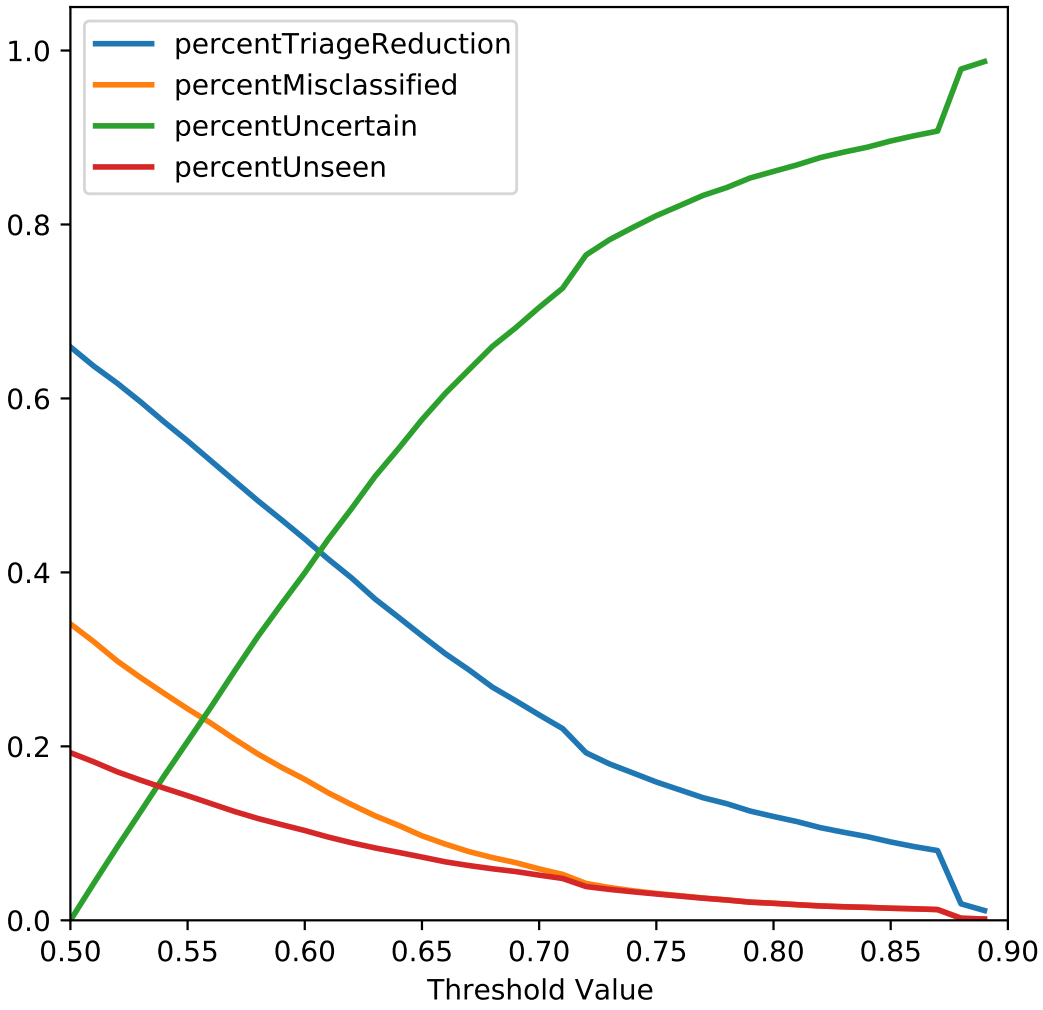

Figure 4.10. triageReduction, percentMisclassified, percentUncertain, and percentUnseen for confidence threshold values ranging from .5 to .9.

systematic bias. The presence of the categories “sexual”, “swearing terms”, and “ridicule” seem to suggest that our false negatives are misclassified not because we bias against a

particular demographic group, but because these misclassified observations contain negative or abusive language. Nevertheless, topical analysis of these false negatives warrants deeper investigation.

4.5.2. Theoretical Improvements and Real World Implications

In practical terms our classifier acts as a relatively effective spam filter for unhelpful feedback. The model filters unhelpful feedback with high precision, but it cannot surface helpful feedback with equal effectiveness. Our topic analysis shows that the model correlates negative emotion, hate, and swearing terms–all topics we could reasonably associate with spam and vandalism–with unhelpful feedback, and that these topics are more explicit in high confidence predictions. Similarly, the bias towards negative predictions at high confidence thresholds indicates that the model filters out spam and vandalism with relatively high precision, but it is less able to predict helpful feedback. As noted above, simply removing spam and vandalism from the feedback queue does result in real work reduction. As illustrated in Table 4.5 and Figure 4.10, we can reduce editor workload with respect to feedback triage at rates of up to 66 percent, depending on our confidence threshold.

However, our classifier's lower performance when making positive predictions indicates that further work would be required to surface or highlight truly insightful reader feedback. Informal spot checking suggests that this lower performance is likely due to the contextual nature of helpful feedback. While spam and vandalism usually contain specific types of vocabulary (e.g. negative emotion, hate, and swearing terms), the utility of helpful feedback depends not only on the text itself, but also on a wide variety of

features related to the article or the editor. For instance, feedback could be more or less helpful depending on the completeness of the article or the type of work the editor prefers to perform. Improving our understanding of helpful feedback and including article or editor features in our featureset represents one possible path towards improving positive feedback predictions.

Even with spam removed from the feedback queue, editors would still need to triage the remaining comments in order to focus their limited time on the most pressing issues. Furthermore, as we work towards the broader goal of bias identification, it is unclear whether this human in the loop triaging process would reintroduce significant editor bias into the reader feedback pipeline. Determining whether our spam filter would make the AFT viable or whether we would need to improve positive classification performance requires additional testing and feedback from the editor community.

4.5.3. Reader-Sourced Knowledge Gap Identification

Our work builds on prior research in several ways. Prior work on peripheral participation [**Lave, 1991**] theorizes that reader-sourcing may reduce barriers to participation and effectively diversify perspectives available to the editor community [**Preece and Shneiderman, 2009**]. However, to our knowledge only one Wikipedia-specific reader-sourced system (the AFT) has been deployed [**Halfaker et al., 2013**], which aimed to collect feedback about low quality articles. Ultimately the AFT was decommissioned after four years of testing and development due to editor push-back, although the Wikipedia Foundation has since developed other successful non-reader-sourced machine learning systems for identifying low quality content, such as ORES [**Halfaker and Geiger, 2020**].

Although the AFT was never framed as a knowledge gap identification tool, findings from our work should be relatively transferable to our broader goal of knowledge gap identification. The AFT aimed to identify and solicit feedback about low quality articles, which represents a subset of all content gaps. A reader-sourced system developed to identify content gaps more broadly would encounter similar challenges as the AFT, specifically lack of adoption due to the extra labor required to sort through large quantities of unhelpful feedback. In this hypothetical system, we could employ a similar filtering and ranking classifier to reduce workload associated with feedback triage. We discuss future work needed to implement a reader-sourced knowledge gap identification pipeline in Chapter 5

CHAPTER 5

Discussion

This dissertation used an asynchronous remote community to investigate Wikipedia editors' existing methods for identifying knowledge gaps and producing content, and trained a machine classifier to identify helpful and unhelpful feedback from Wikipedia's readers. In the following section we explore future directions and limitations for this line of research.

5.1. Non-Reader-Sourced Approaches and Solutions

As illustrated by Study 1, the reasons for inequitable knowledge representation on Wikipedia are both numerous and complex. Reader-sourcing offers one potential solution to one component of the problem–the contributor gap–but ultimately the plurality of barriers calls for a variety of solutions. Although reader sourcing can effectively diversify Wikipedia's contributor base by lowering barriers to entry, there are other issues it cannot solve, such as Wikipedia policy or point of view editing. Additionally, reader-sourcing will always rely on an already time constrained population of expert editors to create content, and while a reader-sourced system can more effectively direct the efforts of these editors it cannot create more person-hours needed to create missing content.

It is beyond the scope of this dissertation to suggest specific interventions for each of these barriers, but one overarching theme from Study 1 suggests that various forms of

outreach and partnerships with experts could provide several paths towards a more equitable encyclopedia. The Wikimedia Foundation relies on professional software engineers, designers, and researches to develop its technical infrastructure, so why not consult experts when improving social and organizational systems? This question reveals one of the fundamental tensions Wikipedia continues to grapple with as it becomes the largest and most reliable single source of knowledge in existence; anyone can participate, yet many tasks–from creating high quality content to updating policy–require expert knowledge. As one editor wrote:

> The notion that good content is a "crowd-sourced," drive-by thing is laughable. There are an active core group of writers, working alone but working collectively, who generate the big part of missing serious content...They need to be identified, treasured, and aided. - p15

To some extent, the Wikimedia Foundation already invests in relationships with external experts. The Wikimedian in Residence program connects external organizations, "typically an art gallery, library, archive, or museum (aka GLAM cultural institutions), learned society, or institute of higher education (such as a university) to facilitate Wikipedia entries related to that institution's mission, encourage and assist it to release material under open licences, and to develop the relationship between the host institution and the Wikimedia community."[1] Some of the Wikimedia Foundation's educational outreach programs use professors and university classes to train future subject matter experts–i.e., students–to participate in Wikipedia.[2] These programs effectively broaden Wikipedia's contributor base, but can be targeted towards individuals and organizations whose expertise focuses

[1] https://en.wikipedia.org/wiki/Wikipedian_in_residence
[2] https://wikiedu.org/

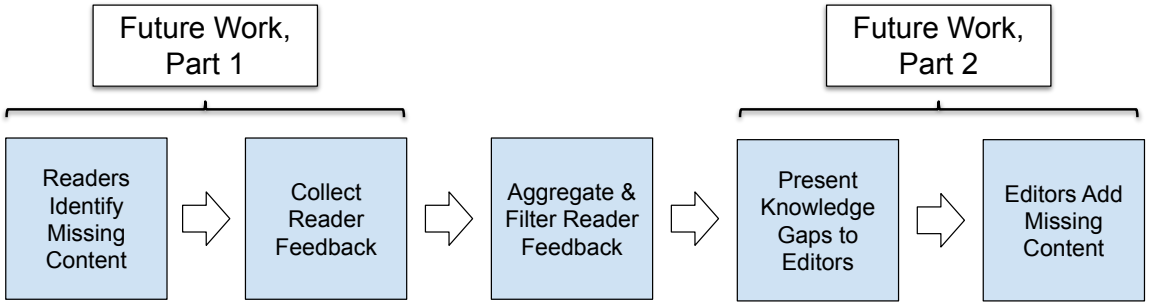

Figure 5.1. Future studies necessary to implement a reader-soured system. Part 1 focuses on the reader facing interface while Part 2 focuses on the editor facing interface.

on underrepresented subjects. Since many core editors already work with outside subject matter experts, continuing to formalize and support these relationships could benefit the encyclopedia. Broadly, while Wikipedia's platform creates novel opportunities and challenges that differ markedly from traditional encyclopedias, not all challenges are novel and not all solutions must be reinvented.

5.2. Future Work for Reader-Sourced Knowledge Gap Identification

Although this work increases the viability of reader-sourced knowledge gap identification, we anticipate several additional steps in order to implement a reader-sourced system. In order to identify knowledge gaps through reader sourcing, we need to: 1) implement and test a reader facing interface and 2) implement and test an editor facing interface. Figure 5.1 illustrates how these two future parts fit into the entire reader-sourced system.

Study 2 classified data collected by the AFT, which solicited readers for suggestions about improving an article's quality. While this dataset contains information about knowledge gaps in the form of low quality content, the tool was never intended to explicitly

identify missing content. Indeed, the AFT's data model, which is built around existing articles, prevents the system from gathering information about subjects that are entirely missing from Wikipedia.

While the AFT's design limits the utility of the existing dataset, a redesign of the reader facing interface to focus explicitly on missing or low quality content provides an opportunity for better overall system performance. The concept of *data scaffolding*, or designing input interfaces and validation systems to ensure high quality data, has been explored extensively in HCI as well as other related domains (e.g. genome sequencing, education, and crowd-sourcing) [**Garg et al., 2019, Díaz and Puente, 2011, Ferretti et al., 2019**]. In future work, researchers could design the reader facing prompts to encourage users to submit higher quality data. For instance, in addition to the open text field, readers could submit additional metadata about missing or low quality content that the classifier could theoretically leverage to improve performance, or that editors could filter by in order to receive feedback targeted for specific editing needs and contexts. A reader might select from a drop-down list of missing content types–such as images, citations, or info-box fields–or directly suggest a missing heading title.

Results from Study 1 reiterate both the challenge discoverability presents for new tool adoption and the need for new tools to integrate with existing workflows. The AFT's initial design used an interface entirely separate from the core workflows editors typically employ, limiting discoverablity and usability.

Fortunately, our participants provided multiple examples of successful tool integration, which we can replicate for an editor facing interface. *Listeria* provides a particularly compelling model, as it generates on-wiki lists of red links, identical to those editors

already create and share by hand. Our reader-sourced system could follow a similar approach and leverage existing technical infrastructure, such as red links, stubs, and various templates, as well as on-wiki completeness lists generated from Wikidata items.

An end-to-end reader-sourced data pipeline for identifying missing content might appear as follows: First, Wikipedia's readers submit scaffolded suggestion data through a reader facing interface, which is designed specifically to collect information about missing content. Second, the system determines whether the missing subject already exists but is incomplete, or whether the subject is missing entirely. Third, using metadata collected in the submission form, the system both determines whether the feedback is likely to be helpful, and routes that feedback to the appropriate WikiProject. Finally, the system adds the subject to the appropriate red link list or the appropriate articles for improvement page, along with a short note explaining the justification for improvement.

Obtaining buy-in from Wikiprojects offers a complementary path towards improving usability and discoverability of any new tool. Again, while the majority of participants in Study 1 were not familiar with Listeria, many had inadvertently used Listeria lists while completing work for Women in Red. Through Women in Red a large population of editors enjoy the benefits of Listeria, but only a small group of administrators are required to interface and use the tool itself. A reader-sourced system for identifying knowledge gaps could use a similar approach.

5.3. Generalizability to Multilingual Wikipedia

while results from Studies 1 and 2 may not easily generalize to other platforms, our findings should translate to other language editions of Wikipedia. Prior work shows that

differences exist across language editions–both in the content that editors create [**Hecht and Gergle, 2010a**] and methods they use to create that content [**Hale, 2014**]–which may require tailored interventions for specific editing cultures. However, our participants in Study 1 were not limited to the English Wikipedia, and many of these editors contributed to multiple language editions. While some of the socio-technical infrastructure (e.g. organized red link lists) may not yet exist in less developed language editions, the methods that editors currently use for identifying knowledge gaps and creating content should be relatively universal, especially as younger language editions mature. Editor communities may need to adapt the identification and production methods explored in Study 1 to fit existing organizational cultures or language edition specific policies, but these changes likely take the form small modifications that update the frameworks outlined in this dissertation.

While the specific models trained in Study 2 would not be able to classify non-English feedback, the methods used to train these models should transfer to other language editions. Indeed, the the Objective Revision Evaluation Service (ORES) uses a similar lexical feature-set in some models, and these models have demonstrated success across a wide range of Wikipedias [**Halfaker and Geiger, 2020**]. The AFT dataset we used in Study 2 contains feedback collected from French and German Wikipedia[3], so future work could use the processing, training, and testing code developed for this dissertation to explore the feasibility of classifying non-English feedback.

[3]In our work we omitted non-English feedback items to simplify the classification task.

CHAPTER 6

Conclusion

This dissertation investigated existing methods for identifying knowledge gaps and producing content, and explored a novel socio-technical system for latent knowledge gap identification. We focus explicitly on Wikipedia's knowledge gaps in order to prioritize concrete and actionable solutions over generalizability. To reiterate, we define a latent knowledge gap as a content gap that is undocumented by prior content gap research and invisible to the editor community. These content gaps represent "unknown unknowns" from the perspective of individuals who produce Wikipedia's content.

In Study 1 (Chapter 3), Wikipedia's editors described their various definitions of "knowledge gaps", their observed causes of those gaps, and their workflows for adding and improving content to fill those gaps. We show that some degree of misalignment exists between editors' definitions and those proposed by the Wikimedia Foundation, but that the causes of gaps observed by editors reinforce findings in prior empirical work. Ultimately, it would likely benefit the long term goals of both editors and the Wikimedia Foundation to reach a consensus about how knowledge gap is defined.

We observe that editors use lightweight markers–e.g. red links, stubs, and templates–in order to distribute labor, and that Wikiprojects leverage these markers to quantify and systematically fill known knowledge gaps. Using aggregation techniques to produce completion lists, Wikiprojects direct attention and labor towards specific topics, even though much of the editor population uses less systematic, serendipitous methods for

choosing which content to work on. The Wikiproject Women in Red represents the most successful and cited example of this model, though others exist on a smaller scale.

Finally, Study 1 reinforces editors' well known resistance to new interventions. With respect to latent knowledge gap identification, many editors expressed frustration towards barriers that exists at other points in the content production process. Editors largely produce content that reflects their interests, and challenges associated with sourcing and writing those articles supersede any need to *identify* more under-produced subjects or underrepresented topics. While developing and employing more sophisticated knowledge gap identification strategies is one important component of creating a more equitable Wikipedia, we cannot effectively fill knowledge gaps without addressing these barriers.

Study 2 (Chapter 4) addressed feasibility issues raised by one novel approach to knowledge gap identification. Reader sourcing offers a possible method for narrowing Wikipedia's contributor gap, but in the past reader-sourced systems have collected a substantial amount of unhelpful feedback, often in the form of malicious commentary or spam.

We trained a machine classifier to differentiate between helpful and unhelpful feedback collected from Wikipedia readers and annotated by editors, achieving an ROC-AUC score of .71. We then explored the possibility of implementing a confidence threshold, or only making predictions for which we achieve a high enough degree of confidence. We show that raising this threshold can result in both relatively high precision as well as a reduction in the amount of triage work required of editors. Our classifier predicts unhelpful feedback with much higher accuracy and confidence than helpful feedback, likely because the factors that make feedback helpful are highly context dependant. By comparison, unhelpful

feedback is generally characterized by similar lexical patterns, such as curse words or hate speech.

Taken together, these studies demonstrate the feasibility of several solutions for identifying latent knowledge gaps. As described in Chapter 5, the relative success of our feedback classifier shows the potential of a reader-sourced feedback pipeline, given that we develop both editor and reader facing interfaces. Additionally, Chapter 3 indicates that other interventions (e.g. policy changes and expanded outreach) must be developed in tandem with technological systems in order to work towards equitable knowledge representation.

Ultimately, inequitable knowledge representation on Wikipedia perpetuates and exacerbates existing inequalities that further disadvantage those "communities that have been left out by structures of power and privilege." [**Zia et al., 2019c**] If Wikipedia is to be a complete compendium of human knowledge, supporting processes and developing systems that can systematically identify and mitigate these knowledge gaps represents a critical step towards democratizing the sum total of *all* human knowledge.

References

[noa, 2021] (2021). Wikipedia Embraces First-of-Its Kind Universal Code of Conduct, Conceived For The New Internet Era.

[Antin and Cheshire, 2010] Antin, J. and Cheshire, C. (2010). Readers Are Not Free-riders: Reading As a Form of Participation on Wikipedia. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 127–130, New York, NY, USA. ACM. event-place: Savannah, Georgia, USA.

[Bao et al., 2012] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. (2012). Omnipedia: Bridging the Wikipedia Language Gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084. ACM.

[Bellomi and Bonato, 2005] Bellomi, F. and Bonato, R. (2005). Network analysis for Wikipedia. In *Proceedings of Wikimania 2005, The First International Wikimedia Conference*.

[Braun and Clarke, 2006] Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

[Bryant et al., 2005] Bryant, S. L., Forte, A., and Bruckman, A. (2005). Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10. ACM.

[Collier and Bear, 2012] Collier, B. and Bear, J. (2012). Conflict, Criticism, or Confidence: An Empirical Examination of the Gender Gap in Wikipedia Contributions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 383–392, New York, NY, USA. ACM. event-place: Seattle, Washington, USA.

[Das et al., 2019] Das, M., Hecht, B., and Gergle, D. (2019). The Gendered Geography of Contributions to OpenStreetMap: Complexities in Self-Focus Bias. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–14, Glasgow, Scotland Uk. ACM Press.

[Díaz and Puente, 2011] Díaz, O. and Puente, G. (2011). Wiki scaffolding: helping organizations to set up wikis. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 154–162, New York, NY, USA. Association for Computing Machinery.

[Elkan, 2001] Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2*, IJCAI'01, pages 973–978, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Ferretti et al., 2019] Ferretti, G., Malandrino, D., Angela Pellegrino, M., Pirozzi, D., Renzi, G., and Scarano, V. (2019). A Non-prescriptive Environment to Scaffold High Quality and Privacy-aware Production of Open Data with AI. In *Proceedings of the 20th Annual International Conference on Digital Government Research*, dg.o 2019, pages 25–34, New York, NY, USA. Association for Computing Machinery.

[Ford et al., 2013] Ford, H., Sen, S., Musicant, D. R., and Miller, N. (2013). Getting to the source: where does Wikipedia get its information from? In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, pages 1–10, New York, NY, USA. Association for Computing Machinery.

[Ford and Wajcman, 2017] Ford, H. and Wajcman, J. (2017). 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social Studies of Science*, 47(4):511–527. Number: 4 Publisher: SAGE Publications.

[Gallert and Van der Velden, 2013] Gallert, P. and Van der Velden, M. (2013). Reliable sources for indigenous knowledge: Dissecting Wikipedia's Catch-22. Accepted: 2013-09-20T13:54:29Z.

[Garg et al., 2019] Garg, K., Kim, Y., Gergle, D., and Zhang, H. (2019). 4X: A Hybrid Approach for Scaffolding Data Collection and Interest in Low-Effort Participatory Sensing. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):90:1–90:28.

[Geiger and Halfaker, 2013] Geiger, R. S. and Halfaker, A. (2013). Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 861–870, New York, NY, USA. Association for Computing Machinery.

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

[Graells-Garrido et al., 2015] Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First Women, Second Sex: Gender Bias in Wikipedia. In *Proceedings of the 26th ACM*

Conference on Hypertext & Social Media, HT '15, pages 165–174, New York, NY, USA. ACM. event-place: Guzelyurt, Northern Cyprus.

[Greenstein and Zhu, 2012] Greenstein, S. and Zhu, F. (2012). Is Wikipedia Biased? *American Economic Review*, 102(3):343–348.

[Greenstein and Zhu, 2016] Greenstein, S. M. and Zhu, F. (2016). Do Experts or Collective Intelligence Write with More Bias? Evidence from Encyclopedia Britannica and Wikipedia.

[Halavais and Lackaff, 2008] Halavais, A. and Lackaff, D. (2008). An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440.

[Hale, 2014] Hale, S. A. (2014). Multilinguals and Wikipedia Editing. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 99–108, New York, NY, USA. ACM. event-place: Bloomington, Indiana, USA.

[Halfaker and Geiger, 2020] Halfaker, A. and Geiger, R. S. (2020). ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):148:1–148:37.

[Halfaker et al., 2013] Halfaker, A., Keyes, O., and Taraborelli, D. (2013). Making Peripheral Participation Legitimate: Reader Engagement Experiments in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 849–860, New York, NY, USA. ACM. event-place: San Antonio, Texas, USA.

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

[He et al., 2018] He, S., Lin, A. Y., Adar, E., and Hecht, B. (2018). The_tower_of_babel.jpg: Diversity of Visual Encyclopedic Knowledge Across Wikipedia Language Editions. In *Twelfth International AAAI Conference on Web and Social Media*.

[Hecht and Gergle, 2009] Hecht, B. and Gergle, D. (2009). Measuring Self-focus Bias in Community-maintained Knowledge Repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies*, C&T '09, pages 11–20, New York, NY, USA. ACM. event-place: University Park, PA, USA.

[Hecht and Gergle, 2010a] Hecht, B. and Gergle, D. (2010a). The Tower of Babel Meets Web 2.0: User-Generated Content and its Applications in a Multilingual Context. In

Proceedings of the SIGCHI conference on human factors in computing systems, pages 291–300. ACM.

[Hecht and Gergle, 2010b] Hecht, B. J. and Gergle, D. (2010b). On the "Localness" of User-generated Content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 229–232, New York, NY, USA. ACM. event-place: Savannah, Georgia, USA.

[Hinnosaar, 2019] Hinnosaar, M. (2019). Gender inequality in new media: Evidence from Wikipedia. *Journal of Economic Behavior & Organization*, 163:262–276.

[Holloway et al., 2007] Holloway, T., Bozicevic, M., and Börner, K. (2007). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, 12(3):30–40.

[Holman Rector, 2008] Holman Rector, L. (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*, 36(1):7–22.

[Hovy et al., 2013] Hovy, E., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

[Johnson et al., 2016] Johnson, I. L., Lin, Y., Li, T. J.-J., Hall, A., Halfaker, A., Schöning, J., and Hecht, B. (2016). Not at Home on the Range: Peer Production and the Urban/Rural Divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 13–25, New York, NY, USA. ACM. event-place: San Jose, California, USA.

[Keegan, 2019] Keegan, B. C. (2019). The Dynamics of Peer-Produced Political Information During the 2016 U.S. Presidential Campaign. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):33:1–33:20.

[Lam and Riedl, 2009] Lam, S. T. K. and Riedl, J. (2009). Is Wikipedia growing a longer tail? In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 105–114, New York, NY, USA. Association for Computing Machinery.

[Lam et al., 2011] Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., and Riedl, J. (2011). WP:clubhouse?: an exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, page 1, Mountain View, California. ACM Press.

[Lave, 1991] Lave, J. (1991). *Situated learning: legitimate peripheral participation*. Learning in doing. Cambridge University Press, Cambridge [England] ; New York.

[Lavin, 2016] Lavin, T. (2016). A Feminist Edit-a-Thon Seeks to Reshape Wikipedia.

[MacLeod et al., 2016] MacLeod, H., Jelen, B., Prabhakar, A., Oehlberg, L., Siek, K., and Connelly, K. (2016). Asynchronous remote communities (ARC) for researching distributed populations. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '16, pages 1–8, Cancun, Mexico. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[Maestre et al., 2018] Maestre, J. F., MacLeod, H., Connelly, C. L., Dunbar, J. C., Beck, J., Siek, K. A., and Shih, P. C. (2018). Defining Through Expansion: Conducting Asynchronous Remote Communities (ARC) Research with Stigmatized Groups. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–13, Montreal QC, Canada. ACM Press.

[Mayfield and Black, 2019] Mayfield, E. and Black, A. W. (2019). Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):206:1–206:26.

[McMahon et al., 2017] McMahon, C., Johnson, I., and Hecht, B. (2017). The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):142–151. Number: 1.

[Metz, 1978] Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. arXiv: 1301.3781.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

[Narayan et al., 2017] Narayan, S., Orlowitz, J., Morgan, J., Hill, B. M., and Shaw, A. (2017). The Wikipedia Adventure: Field Evaluation of an Interactive Tutorial for New Users. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative*

Work and Social Computing, CSCW '17, pages 1785–1799, New York, NY, USA. ACM. event-place: Portland, Oregon, USA.

[Ortega et al., 2008] Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. (2008). On the inequality of contributions to Wikipedia. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 304–304. IEEE.

[Preece and Shneiderman, 2009] Preece, J. and Shneiderman, B. (2009). The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction*, 1(1):13–32.

[Protonotarios et al., 2016] Protonotarios, I., Sarimpei, V., and Otterbacher, J. (2016). Similar Gaps, Different Origins? Women Readers and Editors at Greek Wikipedia. In *Tenth International AAAI Conference on Web and Social Media*.

[Reagle and Rhue, 2011] Reagle, J. and Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, 5(0):21.

[Redi et al., 2021] Redi, M., Gerlach, M., Johnson, I., Morgan, J., and Zia, L. (2021). A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). *arXiv:2008.12314 [cs]*. arXiv: 2008.12314.

[Schneider et al., 2012] Schneider, J., Passant, A., and Decker, S. (2012). Deletion discussions in Wikipedia: decision factors and outcomes. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 1–10, New York, NY, USA. Association for Computing Machinery.

[Sen et al., 2014] Sen, S., Li, T. J.-J., Team, W., and Hecht, B. (2014). WikiBrain: Democratizing Computation on Wikipedia. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, pages 27:1–27:10, New York, NY, USA. ACM.

[Shaw and Hargittai, 2018] Shaw, A. and Hargittai, E. (2018). The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing. *Journal of Communication*, 68(1):143–168.

[Shaw and Hill, 2014] Shaw, A. and Hill, B. M. (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication*, 64(2):215–238.

[Shi et al., 2019] Shi, F., Teplitskiy, M., Duede, E., and Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, 3(4):329–336. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 4 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Business and management;Complex networks;Science, technology and society;Scientific community;Sociology

Subject_term_id: business-and-management;complex-networks;science-technology-and-society;scientific-community;sociology.

[Taraborelli and Ciampaglia, 2010] Taraborelli, D. and Ciampaglia, G. L. (2010). Beyond Notability. Collective Deliberation on Content Inclusion in Wikipedia. In *2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*, pages 122–125.

[Tashman and Edwards, 2011] Tashman, C. S. and Edwards, W. K. (2011). Active reading and its discontents: the situations, problems and ideas of readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2927–2936, New York, NY, USA. Association for Computing Machinery.

[Viegas, 2007] Viegas, F. B. (2007). The visual side of wikipedia. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 85–85.

[Wagner et al., 2015] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *arXiv preprint arXiv:1501.06307*.

[Wagner et al., 2016] Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, 5(1):5.

[Walker and DeVito, 2020] Walker, A. M. and DeVito, M. A. (2020). "'More gay' fits in better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–15, New York, NY, USA. Association for Computing Machinery.

[Warncke-Wang et al., 2013] Warncke-Wang, M., Cosley, D., and Riedl, J. (2013). Tell Me More: An Actionable Quality Model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, pages 8:1–8:10, New York, NY, USA. ACM.

[Warncke-Wang et al., 2015] Warncke-Wang, M., Ranjan, V., Terveen, L., and Hecht, B. (2015). Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. In *Ninth International AAAI Conference on Web and Social Media*.

[Welser et al., 2011] Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. (2011). Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference*, pages 122–129. ACM.

[Wexelbaum et al., 2015] Wexelbaum, R., Herzog, K., and Rasberry, L. (2015). Queering Wikipedia. *Queers Online: LGBT Digital Practices in Libraries, Archives, and Museums*.

[Zia et al., 2019a] Zia, L., Johnson, I., Mansurov, B., Morgan, J., Redi, M., Saez-Trumper, D., and Taraborelli, D. (2019a). Foundations - Wikimedia Research 2030. *Figshare*.

[Zia et al., 2019b] Zia, L., Johnson, I., Mansurov, B., Morgan, J., Redi, M., Saez-Trumper, D., and Taraborelli, D. (2019b). Knowledge Gaps – Wikimedia Research 2030. *Figshare*.

[Zia et al., 2019c] Zia, L., Johnson, I., Mansurov, B., Morgan, J., Redi, M., Saez-Trumper, D., and Taraborelli, D. (2019c). Knowledge Integrity - Wikimedia Research 2030. *Figshare*.

APPENDIX A

List of Recruitment Channels

A.1. Wikipedia and Wikimedia

- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_writers/Missing_articles
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Missing_encyclopedic_articles
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Notability
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Requested_articles
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Intertranswiki
- https://en.wikipedia.org/wiki/Wikipedia:Vaccine_safety
- https://meta.wikimedia.org/wiki/Talk:Gender_Diversity_Visibility_Community_User_Group

A.2. User Groups

- https://meta.wikimedia.org/wiki/WikiBlind_User_Group
- https://meta.wikimedia.org/wiki/Wikimedia_and_Libraries_User_Group
- https://meta.wikimedia.org/wiki/Wikipedia_%26_Education_User_Group
- https://meta.wikimedia.org/wiki/Wikisource_Community_User_Group

- https://meta.wikimedia.org/wiki/WikiWomen%27s_User_Group
- https://meta.wikimedia.org/wiki/Para-Wikimedians_Community_User_Group
- https://meta.wikimedia.org/wiki/Wikimedia_LGBT%2B/Portal

A.3. 3rd Party Organizations

- Whose knowledge: https://whoseknowledge.org/
- Wikiedu: https://wikiedu.org/professional-development/women-in-red/?pk_campaign=women%20in%20red%20-%20social%20science%20-%20jan%2024
- Afrocrowd: https://afrocrowd.org/
- Afrocrowd List of affiliates: https://afrocrowd.org/outreach-partners/
- Black lunch table: https://blacklunchtable.com/

APPENDIX B

Screener

- What is your gender?
  - Male
  - Female
  - Non-Binary
- What is your age?
  - Under 18 (disqualify)
  - 18 to 29
  - 30 to 39
  - 40 to 49
  - 50 to 59
  - Over 60
- What is your household income?
  - Less than $25,000
  - $25,000 - $49,999
  - $50,000 - $74,999
  - $75,000 - $99,999
  - $100,000 or more
- What is your race?
  - American Indian or Alaska Native

  - Asian
  - Black or African American
  - Native Hawaiian or Other Pacific Islander
  - White
- What is the highest degree you have earned?
  - Less than a high school diploma
  - High school degree or equivalent (e.g. GED)
  - Some college, no degree
  - Associate degree (e.g. AA, AS)
  - Bachelor's degree (e.g. BA, BS)
  - Master's degree (e.g. MA, MS, MEd)
  - Doctorate (e.g. PhD, EdD)
- What is your country of residence?
- How often do you edit Wikipedia?
  - Daily
  - Several times a week
  - Several times a month
  - Less than once a month (disqualify)
- In your own words, how would you define a "knowledge gap"?
- Which WikiProjects are you affiliated with?
- Which subject areas or specific articles (if any) do you tend to focus on?
- Which types of work do you tend to accomplish as a Wikipedia editor? How would you briefly describe your role? Please note any work related to finding or

mitigating knowledge gaps. (disqualify if no mention of knowledge gaps, missing knowledge, or bias).

- What is your profession?
- Which languages do you speak fluently? (disqualify if list does not include English)
- Where did you hear about this study?
- Why are you interested in this study? What do you hope to gain from participating in this study?

APPENDIX C

Code of Conduct

Below are copies of The Code of Conduct and the moderation guidelines. We put in place a Code of Conduct for our Asynchronous remote community to limit unacceptable behavior and protect participants from unreasonable harms. Language for the Code of Conduct was adapted for our context from the Working Agreements for Community Cave Chicago, QACON 2019, and prior ARC studies [**Walker and DeVito, 2020**]. Additionally, we created moderation guidelines to ensure equitable enforcement of the Code of Conduct, as well as provide clear structures for escalating reporting of problematic behavior.

C.1. Code of Conduct

Our research group is meant to be a safe and open space for our participants. As such, the group operates with the following code of conduct:

(1) **You Know You, I Know Me:** Make no assumptions about others. When speaking, please try to use "I" statements and avoid making generalizations or applying your own ideals to others.

(2) **What happens here stays here:** Though you are welcome to share your experiences, feelings, etc with others afterwards, please refrain from repeating others' stories, names, likenesses, etc outside of the group.

(3) **Oops/Ouch:** If something offensive, problematic, or hurtful is said or done during group, anyone may say, "ouch." The person that had been speaking should please say, "oops," and then the problematic nature of what was said should be discussed by those persons and/or the group.

(4) **Ouch, Anon:** If any person feels that an "ouch" needs to be said, but is not comfortable saying so at the moment of occurrence, this should be communicated to our moderators. If you are comfortable identifying yourself, use the "Report to Admin" tool on the problematic post, or DM one of the moderators. If you wish to report anonymously, a form which will send an anonymous report to our moderator channel will be provided.

(5) **Know & Check Your Privilege:** Be conscious that all of us view life through our own lens. Our goal is to learn from each other and celebrate our diverse narratives.

(6) **Assume Positive Intent:** Not everyone comes in with the same set of experiences and knowledge, so assume that the folk here are speaking with good intent and the willingness to learn. That being said, hold yourself to the same standard and be accountable for the impact of your actions.

(7) **Don't Giggle My Wiggle:** Folks here have different tastes and preferences, so avoid antagonizing language like "I hate that."

C.2. Moderation Guidelines

If a violation of our code of conduct occurs, we follow a three-level procedure for dealing with incidents:

Level 1: Participants are encouraged to first respond to posts or responses they find problematic by employing the "Oops/Ouch" principle from our working agreements. This is especially true in cases where the intent is clearly not expressly to offend. If you are comfortable, participants are encouraged to post a short response to the comment in question indicating that you would prefer folks to avoid that type of posting and why, then lead the topic gently back in the right direction with some substantive comment on the subject matter in discussion. In cases where offense appears to be the intent, participants are encouraged to escalate to the "Ouch, Anon" principle.

Level 2: In the case of a report from a participant (as laid out in the "Ouch, Anon" principle), or a case of obvious malicious trolling or hate speech, moderators will review the post in question and, if appropriate, record the content of the post for future analysis and remove the original from the thread. The moderator will notify the participant of this privately via direct message and explain how the response is not within the group guidelines, requesting that further responses of that nature not be entered into the group conversation.

Level 3: In the case of repeated violation of our working agreements (e.g., 3 or more incidents), a project co-investigator/administrator will make a decision as to the offending participants continued participation in the research community. This decision will largely be based on the participant's effect on the ongoing safety and norms of openness for the group as a whole. Repeated offenders may be asked to leave the group as a last resort, and only after following the steps outlined in the procedures above have been followed. By the time a participant is banned, it

should have been made very clear to them that they are behaving unacceptably and have been informed of the terms of continued participation before they are banned. Being asked to leave the group will not require the offending participant to forfeit their initial payment for participating in the study, however they will not be allowed the opportunity to participate in the follow-up interviews and subsequent payment.

APPENDIX D

Codebook

D.1. Knowledge Gap Definitions

Bias	content not backed up by sources
	misnaming
	not the majority POV
Gaps	content missing from a topic area
	distorted, biased, or POV content
	from newly published material
	in one language edition but not another
	in other sources but not in WP
	incomplete or missing content
	more articles about one group than another
	notable current events
	omission of one or more narratives
	outdated information
Misc	miscategorization of content
	misgendering
	missing content is a type of bias
	some gaps are more worth filling than others
	there are known underrepresented groups

D.2. Knowledge Gap Causes

Contributor Gap	editors work on what they know
	editors work on what interests them
	lack of time/people
	gatekeeping practices
	low new editor retention
POV Editing	cherry picking sources
	political bias
	conflict of interest
	cultural bias
	editor burnout
	unconscious bias
notability	notability is subjective
	notability is not the cause of information deficit
	notability improves quality
	completeness lists confir notability
sources	source bias/reliability
	paywalls
	lack of sources
	non-western sources
	different language sources

D.3. Knowledge Gap Identification and Content Creation Methods

Passive	failed search on WP
	hard to prove bias
	read external source, search WP
	stumble on bias
	discover missing info while reading WP
Active	patterns of missing content
	systematic focus on underrepresented topics
	use category structure to identify gaps
	on-wiki lists
	watch specific articles for bias
	collaboration with external experts
	wikiprojects document missing articles
	external lists
	active search for known biases
	personal completion lists
	AfD/AfC filtes biased articles
	redlink lists for events
	read AfD
Tools	not-in-the-other-language
	listeria
	petscan
	SPARQL
	python
	google docs
Signposting	request help from wikiproject editors
	article placeholder tool
	redlink signposting
	bias tags
	stubs
Misc	experts know where to find gaps

D.4. Proposed Solutions

Bias Prevention	article ownership
	de-anonymization
Wikiprojects	gap task force
	track systemic bias
Automated Tools	template articles
	biography completeness dataset
	listeria and auto list generation
	suggestbot
Recruitment and Access	offline WP
	paid editing
	Wikipedian in Residence
	edit-a-thons
	newcomer mentoring
	students
Policy Changes	external organizations can change policy
	universal code of conduct
	nuanced notability/sourcing guidelines
	wikiprojects can change policy
External Lists	organizations can produce completeness lists
	school curriculum can be used for events
Misc	expert clearing house
	arbitration
	shared personal lists
	nuanced redlinks/tags
	WP library
	tool discoverability
	reader sourcing
	problems can't be solved with tools

APPENDIX E

List of Prompts

Introductions

(1) For your first post, introduce yourself to the community. Please include any background information you feel comfortable sharing (e.g. cultural background, particular interests or hobbies, your profession, or a fun fact), as well as some information about your experience as a Wikipedian. Questions you might answer in your post include the following:

   (a) When did you start contributing to Wikipedia, and why did you start?

   (b) How frequently do you contribute to Wikipedia? How do you contribute?

   (c) Do you contribute to specific articles or specific types of articles? Do you belong to any WikiProjects?

   (d) As a contributor, do you have a specific role or function within the community?

Missing and Biased Content

(2) For your second post we would like you to think about your personal experience with content that is missing from Wikipedia and walk us through a specific example. Can you think of a specific time when you identified content that was missing from Wikipedia? "Missing content" could be an incomplete article, or an article that does not yet exist. How did you identify missing content in this

example? What was your process? Did you use any tools, lists, or other aids to complete this process? What are the challenging parts of this process or barriers you experienced while adding this content? Was this example a "typical" type of editing you perform on Wikipedia?

(3) Similar to the last post, for your third post we would like you to think about your personal experience with biased content. Can you think of a specific time when you identified biased content on Wikipedia? How do you define or conceptualize biased content? How is biased content different from missing content, if at all? Are there differences and/or similarities between your experience with missing content (which you described in your previous post), and biased content?

(4) For your fourth post, think about other Wikipedia contributors who you may interact, work, or collaborate with. These could be individual contributors, WikiProjects (or other groups), or more formal organizations. Do you interact with other contributors (in a positive or negative way) when identifying missing content or bias? Describe how you collaborate (or alternatively describe a conflict) and use specific examples when possible. Do you know of other Wikipedia contributors or communities that seek missing or biased content on Wikipedia? How are these communities similar or different to your own? If you choose to describe interactions with specific individuals, please refrain from using identifiable names.

(5) For your fifth post we would like you to think broadly about "knowledge gaps". How would you define a "knowledge gap" in the context of Wikipedia? You can reuse or modify the definition you gave in the screener survey. How are the

experiences with missing content and/or bias you discussed in prior responses related to your definition? With this definition in mind, think about the specific challenges you face when addressing a knowledge gap. What is the biggest hurdle or obstacle you encounter? With this definition in mind, think about the specific challenges you face when addressing a knowledge gap. What is the biggest hurdle or obstacle you encounter? If you could implement something to aid this process (a tool, a policy change, a new collaborative effort) what would it be?

Workflow

(6) For your sixth post, think explicitly about your process. Sketch a diagram or flowchart that illustrates the steps you take to identify and fix missing or biased content. Make sure to label each step of the chart so that others can easily understand your routine. If you have multiple processes you can draw multiple diagrams, or focus on the one you find most interesting. Finally, indicate parts of this process that are challenging. Explain these challenges in a few sentences.

(7) For your seventh post, think about solution (this could be community based, a new on or off-wiki tool, or something else entirely) that could help with the challenges in the process you diagrammed in the previous exercise. Where would it fit into your routine? If you do not have a specific routine, what task would it help complete? What problem would it solve? What information would it need, and what information would it produce? Some of you already touched on solutions in your previous post, but try to be as specific as possible and to ground

your solution in your own experience as a contributor. If it's easier to describe your solution visually, feel free to upload a sketch or a diagram.

Wrap Up

(8) Your final post will have two parts. First, are there particular initiatives or tools related to missing or biased content that had potential utility but ultimately failed? Why did they fail? What pitfalls should future efforts avoid? Second, consider the responses you and other participants have provided over the past few weeks. Are there other details, factors, anecdotes, or experiences that we did not ask about? If you could highlight one part of one response as a most important takeaway, which part would you highlight? Thank you all for participating in this study. By the end of next week we will send a follow up email with payment details as well as a few logistical questions. As always, feel free to reach out with questions or comments.