COLLABORATIVE CHALLENGES IN WIKIPEDIA AND WIKIDATA:

STRIKING A BALANCE BETWEEN QUALITY AND QUANTITY

Paramita Das

COLLABORATIVE CHALLENGES IN WIKIPEDIA AND WIKIDATA:

STRIKING A BALANCE BETWEEN QUALITY AND QUANTITY

Thesis submitted to the
Indian Institute of Technology, Kharagpur
For award of the degree

of

Doctor of Philosophy

by

Paramita Das

Under the supervision of

Prof. Animesh Mukherjee

COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

April 2025

APPROVAL OF THE VIVA-VOCE BOARD

Date: 26/03/2025

Certified that the thesis entitled **"Collaborative Challenges in Wikipedia and Wikidata: Striking a Balance Between Quality and Quantity"** submitted by Paramita Das to the Indian Institute of Technology, Kharagpur, for the award of the degree of Doctor of Philosophy has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today.

(Member of DSC) (Member of DSC) (Member of DSC)

(Supervisor)

(External Examiner) (Chairman)

SRINATH
SRINIVASA
21.03.2025

CERTIFICATE

This is to certify that the thesis entitled ***"Collaborative Challenges in Wikipedia and Wikidata: Striking a Balance Between Quality and Quantity"****, submitted by Paramita Das to the Indian Institute of Technology, Kharagpur, for the partial fulfillment of the award of the degree of Doctor of Philosophy in Computer Science and Engineering, is a record of bona fide research work carried out by her under my supervision and guidance. The thesis in my opinion, is worthy of consideration for the award of the degree of Doctor of Philosophy in accordance with the regulations of the Institute. To the best of my knowledge, the results embodied in this thesis have not been submitted to any other University or Institute for the award of any other Degree or Diploma.*

Animesh Mukherjee Date:

Professor

Department of Computer Science and Engineering

IIT Kharagpur

DECLARATION

I certify that

a. The work contained in this thesis is original and has been done by me under the guidance of my supervisor.

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in preparing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving the required details in the references.

Paramita Das

ACKNOWLEDGMENTS

Completing a thesis signifies the end of a challenging and rewarding chapter in a scholar's life. While the thesis is a testament to the academic progress made, the acknowledgments section allows for a heartfelt pause to recognize the invaluable support and encouragement received along the way. I want to use this opportunity to extend my deepest thanks to everyone who has been a part of this journey.

Before anything else, I would like to thank my supervisor, Prof. Animesh Mukherjee, from the Department of Computer Science and Engineering. Oprah Winfrey once said, "A mentor is someone who allows you to see the hope inside yourself." Oprah Winfrey's words resonate deeply when I think of Prof. Mukherjee. Beyond his extraordinary research skills, he has been the guiding force that consistently helped me recognize the potential within myself, even during moments of doubt. Throughout my PhD journey, he has fostered a supportive and encouraging environment, allowing me to pursue my goals with renewed confidence, especially during the challenging phases of paper rejections. Thanks to his guidance, I could publish papers at top-tier international conferences. From problem statement discussions to paper writing, he provided constant assistance and invaluable advice. His belief in my abilities has not only shaped my research but has also inspired me to strive for excellence in all my endeavors during my time at IIT Kharagpur.

I sincerely thank the members of my thesis committee—Prof. Shamik

Sural, Prof. Pawan Goyal, Prof. Sourangshu Bhattacharya, and Prof. Koustav Rudra—for their encouragement and constructive criticism that helped improve this thesis. I also acknowledge the inspiration and encouragement received from other faculty members in the Complex Networks Research Group (CNeRG), namely, Prof. Niloy Ganguly, Prof. Saptarshi Ghosh, Prof. Bivas Mitra and Prof. Mainack Mondal. I would also like to thank Diego Saez-Trumper, the senior research scientist, and Leila Zia, the head of research at the Wikimedia Foundation, for extending an internship opportunity at this prestigious research group. My sincere thanks go to Pablo Aragon and Isaac Johnson, research scientists at the Wikimedia Foundation, for their invaluable help in publishing my research work from this internship in a top-rated conference. My sincere gratitude also goes out to Deepak, Dolu, Haimanti-di, and Bappa-da for their immense support in handling official procedures.

I have been extremely fortunate to have found mentors and friends of all kinds in the CNeRG group. I want to extend deep regards to Dr. Soumay Sarkar (Soumya-da) for being my mentor during the early days of my PhD and for his constant involvement in discussions to help me achieve my research goals. Soumya-da has inspired me greatly with his poise and his remarkable ability to handle tricky situations wonderfully. I would like to thank Binny Mathew, Rima Hazra, Abhishek Dash, Soumi Das, Souvic Chakraborty, Bishakh Ghosh, Sayantan Adak, Punyojay Saha, Rajdeep Mukherjee, Bishal Santra, Anurag Roy, Sidharth Jaiswal, and Kiran Purohit. I appreciate the interesting conversations and frequent coffee breaks with them. I would also like to thank my team members, Bhanu Prakash Reddy Guda, Sasi Bhushan Seelaboyina, Anirban Panda, Sai Keerthana Karnam, Debajit Chakraborty, Aditya Soni, Amartya Roy, and

Ritabrata Chakraborty, for their unwavering support throughout my PhD.

I would also like to express my gratitude to Mrs. Soumita Mukherjee for providing homely support away from home. My warm thanks go to Biswajit Sethi, Ipsita Koley, Eashita Chowdhury, Riya Sadhukhan, and Amrita Mondal for constant support during my PhD days.

I extend my heartfelt gratitude to my childhood teachers– Shyamal Dutta and Mahaprasad Dutta, for instilling in me the aspiration to pursue a Ph.D. I am also deeply thankful to Prof. Abhik Mukherjee, Professor in the Department of Computer Science and Engineering at IIEST Shibpur, whose guidance during my M.Tech days helped shape my path toward research.

Last but not least, I would like to acknowledge the constants in my life who have been there for me unconditionally—- Maa, Baba, Joy Mama, Somnath Mama, Kaku, Kakima, and my childhood friends– Madhu, Adwi, and Shilu. Thank you, Subhadip, for being my philosophical mirror and inspiring and supporting me in ways that words cannot fully capture. None of this would have been possible without your love, encouragement, appreciation, support, and constructive criticism. Finally, I would like to thank the Almighty for supporting me through every phase of this wonderful journey.

Paramita Das

IIT Kharagpur, India

Author's Biography

Paramita Das received her B. Tech. degree in Computer Science and Engineering from Academy of Technology Adisaptagram (WBUT), India in 2015. She pursued her M.Tech degree in Computer Science and Engineering from Indian Institute of Technology in Engineering Science & Technology (IIEST) Shibpur in 2018. She obtained PhD degree from the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India, in 2025. Her research interest lies in Natural Language Processing, Deep Learning, Information Retrieval, and Computational Social Science.

Publications from the Thesis

1. **Paramita Das**, Bhanu Prakash Reddy Guda, Sasi Bhushan Seelaboyina, Soumya Sarkar and Animesh Mukherjee. *"Quality Change: Norm or Exception? Measurement, Analysis, and Detection of Quality Change in Wikipedia"*. Proceedings of the ACM on Human-Computer Interaction (2022), Volume 6, Issue CSCW1, Article No. 112, pp. 1 – 36, DOI: https://doi.org/10.1145/3512959
2. **Paramita Das**, Amartya Roy, Ritabrata Chakraborty and Animesh Mukherjee. *"On the effective transfer of knowledge from English to Hindi Wikipedia"*. Proceedings of the 31st International Conference on Computational Linguistics: Industry Track (2025), pp. 453-465, DOI: https://aclanthology.org/2025.coling-industry.39/
3. **Paramita Das**, Sai Keerthana Karnam, Anirban Panda, Bhanu Prakash Reddy Guda, Soumya Sarkar and Animesh Mukherjee. *"Diversity matters: Robustness of bias measurements in Wikidata."* Proceedings of the 15th ACM Web Science Conference (2023), pp. 208 – 218, DOI: https://doi.org/10.1145/3578503.3583620

4. **Paramita Das**, Sai Keerthana Karnam, Aditya Soni, and Animesh Mukherjee. *"Social Biases in Knowledge Representations of Wikidata separates Global North from Global South."* Proceedings of the 17th ACM Web Science Conference (2025).
DOI: https://doi.org/10.1145/ 3717867.3717882

ABSTRACT

Collaborative platforms and their significant contributions to fostering the continuous evolution of knowledge while transcending traditional geo-social boundaries have become a common paradigm in today's digital world. Despite advancements in knowledge democratization, challenges remain in maintaining the quality standards of the vast amounts of content produced. The underlying solution is to develop system-specific frameworks that monitor various quality aspects, especially reducing barriers in sharing knowledge around the globe. In this thesis, we focus on the collaborative challenges of two popular open-access platforms: Wikipedia and Wikidata. We aim to explore different challenges arising from the open-editing model and propose AI-driven automated solutions to address the challenges in these two knowledge hubs.

As the first objective, we study the evolution of the quality of Wikipedia articles from their inception and identify different patterns in quality fluctuations over time. In addition, we propose a novel framework for predicting such dynamic changes in quality. As the second objective, we aim to address content disparity among different language versions of Wikipedia on specific topics, focusing on bridging the knowledge gap between high-resource and low-resource languages. As the third objective, we investigate social biases in Wikidata related to sensitive attributes such as gender and age. The first among these is to identify the diversification of social biases in the Wikidata knowledge graph that persists in the form of data bias and algorithm bias. The second is to understand the impact of these biases on fairness in downstream tasks like link prediction.

Keywords: collaborative platforms, Wikipedia, Wikidata, article quality,

change point detection, knowledge equality, machine translation, knowledge graph, social biases, fairness, link prediction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's digital age, knowledge generation, sharing, and refinement transcends traditional boundaries, giving rise to *collaborative knowledge generation systems*, such as Stack Overflow, Quora, OpenStreetMap, Apache Software Foundation, etc. These decentralized approaches leverage the collective intelligence of a diverse and globally dispersed group of individuals, integrated by digital platforms and open-access resources. As a result, millions of people, separated by thousands of miles can collaborate on a single project. A quintessential example of this paradigm is Wikipedia, an online encyclopedia maintained by volunteers worldwide. Another promising example is Wikidata, a robust, collaborative, and open-source knowledge base that serves as a central repository for structured data across the Wikimedia projects, as well as a myriad of other applications and services. Wikipedia's open-editing model permits anyone with Internet access to contribute and enhance articles, creating a continuously evolving and expanding repository of knowledge. This democratization of knowledge facilitates the exchange of diverse ideas, perspectives, and information, ensuring a more comprehensive and nuanced understanding of various topics. This openness fosters greater transparency, allowing for widespread review, critique, and improvement of the knowledge as required. Further, Wikipedia's extensive content is valuable for numerous applications, including academic research, website and blog content integration, and software development.

Wikipedia articles are also instrumental in advancing NLP tasks, such as training large language models (LLMs), which are crucial for advancements in language understanding, machine translation, and information retrieval. Given Wikipedia's pivotal role in the accumulation and dissemination of knowledge, it employs a robust ecosystem of checks and balances to ensure the accuracy and reliability of its content. We focus on studying several critical facets of this ecosystem to enhance automation in monitoring and managing this expansive knowledge generation system.

Wikipedia moderators ensure that the content of every article should be processed through the collective system [196] and the mechanisms include a community-driven peer-review process where editors monitor changes, verify sources, and uphold content standards, aka *article quality* through guidelines and policies. The Wikipedia community adheres to a detailed set of guidelines to enrich content without compromising standards. Based on these guidelines, articles are categorized into several quality classes by the editors, with increasing adherence to the standards. The hierarchy and categorization of quality classes differ across different languages of Wikipedia [52]. However, this quality assessment task is laborious and demands platform expertise. In recent years, the Wikimedia Foundation has sought automated tools, for example, ORES [1] for evaluating the quality of Wikipedia articles. Further, researchers develop several AI-based solutions (including state-of-the-art machine learning and deep learning approaches) to measure the quality of an article at a specific timestamp by leveraging the diverse collaborative features of Wikipedia as well as the article content features [73, 190]. These prediction models can eliminate the time lag associated with manual efforts but may introduce noise in measuring the dynamic changes of articles. The quality of Wikipedia articles is not static; it changes with modifications to the existing content. In this context, an intriguing question is how Wikipedia articles progress through different quality states over time. Further, the existing quality prediction models are not effective in predicting the dynamic changes in quality.

Besides its openness, Wikipedia remains the largest web-based encyclopedia, encom-

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

passing over 63.2 million articles across 331 languages[2]. Its multilingual content makes Wikipedia a valuable resource for various NLP applications, including multilingual and cross-lingual information retrieval, question-answering systems, etc. However, there is a significant knowledge gap across different language editions of Wikipedia [62], leading to an information divide among users of various language versions. The English Wikipedia is the largest, with over 6.8 million articles, while other language editions, even for widely spoken languages, have considerably fewer articles. For instance, the Hindi Wikipedia has only 162,007 articles as of June 2024, despite Hindi being the third most spoken language globally and the most spoken language in India with ~ 528.3 million speakers. Also, several research studies highlight the information disparity in different language versions of Wikipedia, revealing substantial variations in content quality, coverage, and completeness that affect the accessibility and comprehensiveness of knowledge across the globe. In their work [118], the authors measured the quality and importance of Wikipedia articles across different languages, revealing significant disparities in the quality of information available in various language editions. Another study quantifies the extent of American-centric content across different language versions of Wikipedia and has demonstrated how cultural and regional biases influence the representation of information in Wikipedia articles [109]. In another work [168], authors examined Wikipedia articles on the histories of all UN member states across 30 language editions. They identified recency bias in Wikipedia narratives that favoring recent events over distant ones. Further, they found that the distribution of historical focal points varied across continents, with similar patterns often aligning with geopolitical blocs. Alongside the cultural differences and relevance of a topic to a particular language community, this content disparity can often stem from the unavailability of active editors in language groups and their editing behaviors. Researchers have shown how different levels of engagement and interest among editors contribute to variations in content quality and completeness across languages [102]. It has been noticed that articles in major languages like English often contain more detailed and up-to-date information compared to those in less widely spoken languages. Consequently, users who rely on non-English versions of Wikipedia may encounter gaps in information or lack access to critical knowledge,

[2]https://en.wikipedia.org/wiki/List_of_Wikipedias

thereby perpetuating educational and informational inequities. The Wikimedia Foundation quantifies this knowledge gap across its projects, including Wikipedia, in a taxonomy format, in which the reduction of content gaps is a major focus [163]. Research efforts are required to bridge this gap including encouraging multilingual contributions and automated machine translation of key concepts, but the challenge remains substantial.

Similar to many other open-sourced *knowledge graphs* (KGs), such as DBpedia, YAGO, OpenCyc, ConceptNet, Bio2RDF, etc., Wikidata is a free and open knowledge base of the structured data that can be read and edited by both humans and machines. Wikidata allows users to create, update, and query data entries, making it a dynamic and continuously evolving resource that acts as a central storage of many Wikimedia projects, such as Wikipedia, Wikivoyage, Wiktionary, Wikisource, as well as other external websites and applications. However, this openness, while fostering a rich and diverse dataset, also introduces the potential for bias [172, 205]. For instance, certain professions in Wikidata may be underrepresented or described in ways that reinforce gender stereotypes. Considering gender as a sensitive attribute, it might well be the case that the profession of automobile racing driver is male-dominated in Middle Eastern countries while it is a gender-neutral occupation in the Western world. Unfortunately, while high-quality structured content is a plus, a wide range of societal and human biases are inherent to KGs in many ways – either in the form of sampling strategy or the judgmental view. A pertinent question in the research community exists regarding the source of biases in knowledge graphs. Researchers have shown that the entities and relationships in a typical KG are accumulated in a (semi) automatic way [55], which may result in gathering biased knowledge from the implicit biases of individuals involved in curating the knowledge graphs and the open text corpus of the web. Further, algorithms used to sample, aggregate, and process knowledge can incorporate biases into KGs. In handshake with the proliferation in embedding learning methods [93], recent works have established the anecdotal presence of societal biases in KG data and how they are being mirrored by state-of-the-art KG embedding algorithms [63]. Biases encoded in KGs and knowledge graph embeddings (KGEs) harm society as well as the underlying automation systems that leverage the knowledge extracted from KGs in building downstream applications. To tackle this issue, researchers have come up with coherent frameworks of bias mea-

surement and debiasing them further [17]. Cultural worldviews significantly affect how bias is defined, interpreted, and addressed. Studies across disciplines—psychology [95], computational linguistics [57, 110], and history [37]—demonstrate that biases are deeply embedded within and influenced by cultural contexts. This necessitates culturally sensitive frameworks when analyzing or mitigating bias in any domain. However, there is a lack of research that focuses on the biases incurred from a sensitive attribute varying across socio-economic, socio-cultural, and geographical boundaries as outlined in the knowledge graphs. It can be the case in which two key design factors– (i) the choice of geo-social data from different geographies of knowledge graphs and (ii) KG embedding representation algorithms –have a significant influence on the behavior of bias measurement in KGs. This is often overlooked by most coarse-grained approaches working at the aggregate level, therefore highlighting the need for a comprehensive audit of bias measurement methods in knowledge graphs (KGs), particularly in large-scale open-source KGs like Wikidata. Apart from the encoded biases within the KGs, biases have a significant impact on downstream applications, specifically link prediction (LP) which helps to address the problem of the incompleteness of the knowledge graphs. In this context, it is worth mentioning that KGs indeed face challenges when it comes to balancing bias versus fairness. As a result, link prediction can lead to unfair prediction of graph links for minority groups. The immediate research question is how social biases affect the downstream link prediction task in predicting an observable, e.g., occupation for different groups categorized in terms of different sensitive attributes, for example– male/female in case of gender, young/old in case of age of a person. Therefore, the biases embedded in KGs can lead to unfair link prediction, particularly affecting minority groups. Similar to the approaches in measuring biases encoded in KGs, the framework for measuring unfairness in link prediction can be further explored based on the design choices– (i) data subset, and (ii) embedding learning algorithms. Understanding the variation of social biases across geographies can shed light on the global variation of geo-social and economic attributes.

1.1 Contributions

We present the salient contributions we have made in this thesis to fulfill the objectives discussed above.

1.1.1 Assessment of quality change in Wikipedia articles

Our key contribution toward understanding the evolution of quality followed by detecting quality change in Wikipedia articles is as follows.

- In this work, we tackle two significant issues regarding the article quality of English Wikipedia articles – (i) evolution of quality throughout the life-cycle of articles, and (ii) prediction of quality change.
- In our study, we conduct extensive experiments to understand the intricate details of quality changes in Wikipedia articles throughout their life cycles. We compile a dataset of approximately 30,000 articles, each with quality assessments by Wikipedia editors for every revision from their creation until June 2019. By analyzing these articles individually, we identify the temporal patterns of their quality evolution, which we term as the article quality life-cycle. We categorize these patterns into four main groups: (a) only promotion, (b) only demotion, (c) both promotion and demotion, and (d) no change in quality. Notably, our 14-year longitudinal analysis (2006-2019) revealed that sequential quality changes are more of an exception than a norm. For instance, 51.73% of articles did not undergo any quality changes after their initial assessment, a trend that has become increasingly common. We also find that lower-quality articles (B, C, Start, Stub) are 64.61% more likely to stagnate than higher-quality articles (FA, GA, A). In addition, we discover a unique type of cyclic shift, where article quality oscillates between promotions and demotions, often returning to its original state. These cyclic shifts, which exhibit short turnaround times (minimum is less than 15 days),

suggest a *quality switch war* similar to *edit wars* [179], likely due to ongoing conflicts among editors over article quality assessment.

- In the second task, we apply state-of-the-art multivariate change point detection algorithms [184] to identify quality changes in Wikipedia articles, treating this as a change point detection problem. We utilize an array of features derived from editor attributes, article characteristics, and edit patterns and achieve coverage of 76% in detecting quality change points. Unlike traditional quality prediction models that use features from multiple pages, our unsupervised, page-level approach focuses on identifying quality changes based on intuitive features specific to each article. This method is simpler and more efficient, avoiding the pitfalls of black-box quality prediction models that often fail to capture the dynamic changes in individual articles. Moreover, our prediction model wins over the existing machine-learning-based quality prediction framework of the Wikimedia Foundation, ORES, by a significant margin (our method outperforms ORES by 21% in coverage and 38% in precision.

1.1.2 Knowledge equity across multilingual Wikipedia articles

Our key contributions are as follows.

- To address the disparity in content quality between Hindi and English Wikipedia articles, we gather approximately 20,000 existing articles in both languages, along with their manual quality assessments from the English versions. Leveraging a quality scoring dataset [52], which assigns scores between 0 and 1 (with lower scores indicating poorer quality), the quality of each article is extracted for every language version individually. We focus on the articles in which the Hindi version's quality score is lower than that of the English version. This results in a subset of around 18,000 articles for further analysis.
- For English articles rated as high-quality (FA), we utilize a two-stage process to enhance their Hindi counterparts. First, we map the sections of the Hindi and En-

glish versions by assessing the semantic similarity of their section headings. Once the sections are matched, the content from the corresponding English sections is machine-translated into Hindi, making it available for integration into the Hindi articles.

- On the other hand, for English articles categorized as GA, A, B, C, Start, or Stub, our framework first enhances their content using external resources before incorporating this improved content into their Hindi counterparts. We source at least one biography written in English from publicly accessible repositories for each article. To extract information from these external biographies aligned with the content of the articles, we employ the standard RAG method. We then ensure the retrieved content complies with Wikipedia's NPOV policy by identifying and removing subjective biases from sentences extracted from the biographies. This is achieved using an in-context learning setup, employing `Llama3` (70B) to generate neutral content from the extracted text. Once the content is neutralized, it is integrated into the Hindi articles through the two stages previously described: section mapping and machine translation.

1.1.3 Auditing bias and fairness in Wikidata

Our key contributions in surveying social biases embedded in Wikidata and their further implications in downstream applications, especially in link prediction, are as follows.

Finding societal biases in Wikidata

- First, we curate a comprehensive dataset comprising 2.22 million Wikidata entities and 894 distinct relations, spanning 13 demographics worldwide, including Arabia, Australia, Argentina, Brazil, France, Germany, India, Japan, Kenya, Russia, South Africa, the United Kingdom, and the United States. This diverse selection ensures representation from all continents. Using these entities, we construct a giant network by connecting edges among them. Here, edges refer to the common relations

between the entities. For our analysis, we focus on gender as a sensitive attribute and limit our examination to the binary gender categories – male and female.

- We investigate the influence of gender as a sensitive attribute on professions within our knowledge graph (KG) dataset in two primary ways: data bias and algorithm bias. First, we assess the existing biases in the dataset (data bias) by implementing the method proposed by researchers [33] to compute a bias score for each profession within a specific demographic. We categorize professions into male-biased, female-biased, and neutral categories using this bias score. Next, we evaluate algorithm bias by generating embeddings from scratch using two graph embedding models, TRANSE and COMPLEX, applied to our demographic dataset. This process follows a bias measurement metric from the previous study [63] but is applied individually to different demographics in our dataset. Each profession in these demographics is assigned a bias score, leading to the creation of ranked lists (in descending order of bias score) of male- and female-biased professions. We then analyze these rankings based on two key dimensions: the embedding learning methods used and the different demographics considered.

- Our large-scale data-driven analysis reveals that the inherent data bias in the dataset is indeed manipulated by the algorithmic bias introduced by embedding learning algorithms, a trend observed across all demographics. The results further indicate that if one discounts the inherent data bias, the embedding learning algorithms themselves introduce biases that can potentially affect the downstream NLP applications that rely on KG embeddings. Next, we compare the biases introduced by the two embedding algorithms. TRANSE highlights generic professions across demographics, whereas COMPLEX ranks demography-specific professions higher. When comparing the rankings of biased professions across the 13 demographics, TRANSE often groups culturally or socio-economically similar regions, such as (a) the UK and the USA or (b) Australia and the UK, despite geographical differences. This suggests that TRANSE's focus on generic professions naturally reflects broader socio-economic and cultural similarities. Conversely, COMPLEX identifies similarities in geographically closer regions, highlighting more nuanced

and specific biases. In addition, we analyze the entropy of occurrence of the top 50 male and female-biased professions across demographics. We find that male-biased professions exhibit greater variety in their distribution across different demographics compared to female-biased professions for both embedding methods. This suggests a broader range of male-biased roles globally, while female-biased roles tend to be more consistent across regions.

Finding bias and fairness in link prediction

- In this work, we expand the previous dataset to have knowledge graph triples from a total of 21 different geographies and associated with two sensitive human entity information, gender, and age. This dataset, sourced from Wikidata, includes 3.2 million human entities along with their occupational details, meticulously organized to support the downstream task of link prediction and to evaluate societal biases in this task. The dataset's uniqueness stems from two main features –

    1. **Geographical categorization**: It contains distinct Wikidata triples for 21 regions worldwide, including Arabia, India, Israel, Japan, South Korea, Turkey, Russia, Australia, New Zealand, Egypt, Nigeria, South Africa, France, Germany, Spain, the United Kingdom, Argentina, Brazil, Canada, Mexico, and the United States. This categorization is based on the country of citizenship of the human entities.
    2. **Sensitive attribute**: Each triple includes details on the gender and age of the human entities. Age is derived from the reported date of birth.

- Next, we align our proposed dataset in seeking answers to the question – how do social biases (i.e., gender: male/female and age: young/old in our work) impact predictive outcomes related to an observable (occupation or profession in our case) in the link prediction task given a knowledge graph. To investigate how social biases affect the fair link prediction of occupations, we propose a novel framework named AUDITLP to measure the biased predictions given the sensitive attributes of human entities. We benchmark our dataset with the help of AUDITLP for three popular

knowledge graph embedding learning (KGE) algorithms – TRANSE, DISTMULT, COMPGCN and GEKC which generate knowledge graph embeddings as the features to be used in the link prediction pipeline. The predictions generated by the framework are analyzed using popular fairness metrics, resulting in the following key findings -– (1) the classifier outcomes are unfair in terms of sensitive attributes for a given set of professions, (2) our large-scale qualitative analysis reveals that the choice of data subset, i.e., geographies, significantly impact the variance of biases. Specifically, the social biases present in different geographies manifest in a clear partition of the world into two distinct regions: the Global North and the Global South, characterized by their different geo-social and economic attributes. Surprisingly, this result is true for all the algorithms we used in AUDITLP – TRANSE, DISTMUSLT, COMPGCN and GEKC – even though their inner workings are quite different. Such a result indicates that this observation has a universal underpinning.

1.2 Objectives

This section outlines the objectives of the thesis that we set forth to solve in this thesis, mainly anchoring on the issues mentioned above. In the following, we succinctly articulate these objectives.

1.2.1 Ecosystem of quality in Wikipedia

In this part of the thesis, we aim to study the evolution of quality scales of English Wikipedia articles which include FA (*featured article* that is professional, outstanding, and thorough), A (the article well organized and complete), GA (*good article* that is useful to nearly all readers but not as good as professional ones), B (the article is mostly complete), C (the article is substantial but not complete), START (the article is still developing and quite incomplete), and STUB (the article has very little meaningful content). Specifically, we plan to investigate how Wikipedia articles change through various

quality classes over time, as the quality often shifts due to ongoing content modifications by the editors. As a secondary objective, we aim to determine the appropriate method for detecting dynamic changes in article quality. We attempt to create an automated, data-driven approach to identify early indicators influencing quality shifts in Wikipedia articles. Further, we are interested in highlighting the most distinctive features of an article that can help predict future changes in its quality.

1.2.2 Ecosystem of knowledge equity in Wikipedia

In this part of the thesis, we aim to address the content disparity in Wikipedia articles between high and low-resourced language versions, specifically English as the high-resourced language and Hindi as the relatively low-resourced one. We target to develop a framework for effectively transferring knowledge from enriched English articles to their less enriched Hindi counterparts on the same topics. To achieve this, we first investigate whether existing content from English Wikipedia can be directly included in Hindi articles or if additional information from an external English web corpus must be integrated with the English articles before transferring this knowledge to Hindi articles. When integrating external content, it is crucial to address various biases, such as framing and epistemological biases, that may conflict with Wikipedia's neutral point of view (NPOV) policy. Therefore, we propose designing a sub-module to adapt external content to align with Wikipedia's NPOV standards. Furthermore, we explore the use of state-of-the-art machine translation techniques to generate suitable, Wikipedia-style Hindi content, which can be further incorporated into Hindi articles, thereby attaining knowledge parity between high-resource and low-resource articles.

1.2.3 Ecosystem of bias and fairness in Wikidata

In this part of the thesis, our objective is to demonstrate the variance of societal biases in Wikidata and its effects on important downstream applications. Our objectives are dual.

As the first objective, we aim to experiment with the biases embedded in Wikidata, aka data bias. Moreover, algorithmic biases surface when knowledge triples are converted into embedding by embedding learning algorithms. Toward this objective, we plan to examine how two key design choices influence bias measurement in knowledge graphs (KGs): (i) the selection of geo-social data, i.e., knowledge graphs related to different geographical regions, and (ii) the choice of KG embedding representation algorithms. In this context, the pertinent question is how these choices impact the variability of biases, for a given sensitive attribute. The base dataset we consider for our data-driven analysis might have biases resulting from cultural differences (aka data bias) belonging to different geographies. Our point of interest is to check whether algorithmic biases creep in when we build embedding from this base dataset. Finally, to achieve this objective, we wish to perform a data-driven audit to understand the variance of gender bias (based on gender division - male and female) across two important orthogonal axes as previously mentioned- (i) choice of training data modality and (ii) embedding learning algorithm. As the second objective, we aim to explore the interplay between bias and fairness in the context of the downstream link prediction (LP) task within the Wikidata ecosystem. Similar to our first objective, we will analyze the impact of two critical design choices —- geo-social data selection and embedding learning algorithms —- on fairness in LP. To achieve this objective, we first intend to create a meticulously curated dataset that includes human entities from 21 different geographies worldwide. Next, we plan to establish a framework for identifying biased outcomes in LP, focusing on how occupations are classified as either male- or female-dominated when gender is considered a sensitive attribute and as young- or old-biased, or age-neutral when age is the sensitive attribute. In addition, we aim to examine how these biased outcomes vary across different geographies, reflecting the socio-economic and cultural divisions of the world. Through this analysis, we seek to provide a deeper understanding of fairness in LP and how biases manifest and propagate across diverse global contexts.

1.3 Organization of the thesis

In this section, we summarize the organization of the rest of the thesis.

- **Chapter 2** describes the related works in different areas of research that are relevant to the objectives of this thesis.
- **Chapter 3** discusses the contribution made toward our first objective. Specifically, we show the evolution quality scale of an English Wikipedia article identifying non-intuitive patterns. Further, we attempt to develop an automated data-driven approach employing state-of-the-art change point detection algorithms for the detection of the early signals influencing the quality change of articles.
- **Chapter 4** discusses the second objective, i.e., addressing the content disparity between multilingual Wikipedia articles. We develop an automated approach to transfer knowledge from a high-resource language like English to a low-resource language like Hindi. This involves enhancing the content of Hindi Wikipedia articles by leveraging both the rich content available in English Wikipedia and additional resources from external English-language web corpus.
- **Chapter 5** sheds light on societal biases that are encoded in knowledge graphs (Wikidata, in our case) and fairness associated with downstream link prediction tasks. Our focus is on understanding how biases and fairness vary across different design axes—choice of geo-social data and embedding learning algorithms—and their implications for socio-economic and cultural divisions globally.
- **Chapter 6** concludes the thesis by reviewing the contributions and outlining possible future efforts that this thesis opens up for the research community.

Chapter 2

Related Work

In this chapter, we discuss the related research works pertinent to the thesis. We list the relevant works denoting various challenges present across collaborative platforms, particularly Wikipedia and Wikidata. Among these challenges, first, we focus on research dedicated to the quality standards of Wikipedia articles and the development of frameworks for automatic article quality assessment. Second, we explore several issues related to knowledge inequality among different language versions of Wikipedia. Lastly, we review studies that measure social biases at different stages of knowledge graphs, with a special emphasis on Wikidata.

2.1 Quality changes in Wikipedia articles

A series of studies revealed that the dynamics of collaboration and diverse set of collaborative features contribute to the creation of high quality content. This topic is well studied in case of Wikipedia, where typical organizational constraints (e.g., protection level of pages - semi-protected, unprotected etc.), editors' authority (e.g., admin, auto-confirmed users etc.), norms and guidelines (e.g., NPOV, three revert rules etc.), technical features, such as bots etc. play key roles behind the individual contribution as well as the quality of

the end product. Longitudinal analysis of Wikipedia articles have resulted in discovering several key factors which could potentially have either positive or negative impact on the collaborative platform [204].

2.1.1 Content quality assessment in Wikipedia

Automatic article quality assessment in Wikipedia involves utilizing heuristics or different machine learning (ML) and deep learning (DL) models to evaluate and predict the quality of articles based on a collection of features and metrics. Automatic article assessment has been explored by both researchers at Wikimedia foundation [75] and academia [25].
Quaity assessment methods: Early research examined the connection between the length of Wikipedia articles and their perceived quality [28]. The authors found that longer articles often offer more comprehensive coverage of a topic, include more references, and have better organization – i.e., all essential components of high-quality Wikipedia entries. Further studies have explored the impact of writing style on identifying high-quality articles, using automated tools to differentiate featured articles from others based on stylistic characteristics [125]. Building on this line of inquiry, researchers have begun to pinpoint various structural features that could effectively predict article quality. For instance, the authors of [195] proposed an actionable framework that evaluates article quality based on several key aspects, including textual quality, structural organization, and readability metrics. This model aims to provide specific, actionable feedback to editors to facilitate targeted improvements. Another study [10] focused on using machine learning techniques to analyze a diverse range of features – e.g., linguistic characteristics, edit history, contributor diversity, and metadata – to identify common issues that may detract the content quality. By incorporating structural features and implementing machine learning models, various studies have suggested models to streamline the often time-consuming process of manual feature engineering [160, 175, 176]. In addition, researchers [53, 123] have explored the relationship between article quality and the structural properties of co-editor and editor-article networks. With the advent of deep learning approaches, several researchers have employed various deep learning models – convolutional neural networks

(CNNs), long short-term memory (LSTM) networks, CNN-LSTMs, and bidirectional LSTMs – to evaluate the quality of Wikipedia articles [49, 73, 189]. It has been observed that different language cultures do not always share the same understanding of quality, which is crucial to consider when designing a multilingual quality assessment solution. Various researchers have addressed this issue using cross-lingual NLP approaches. In their works [116, 117], the authors identified measures for assessing the quality of data in the structural parts of Wikipedia articles—such as infoboxes across multiple languages. They suggested that selecting the best language versions of a particular article can enrich less developed versions of the same article. Another research provides a comprehensive evaluation of Wikipedia articles in different languages, using automated tools to assess the relative quality and popularity of articles [120]. The study [122], which focused on multilingual quality aspects, introduced a transfer learning approach to detect quality flaws in Wikipedia articles. This research involved building a cross-language classification model using transfer learning to identify flaws, specifically detecting advert flaws (content written like an advertisement) in French, Spanish, and Chinese articles, with English articles serving as the source language. Their experimental results demonstrated that multilingual BERT, trained with the English dataset, could identify advert flaws in other languages, and that fine-tuning transfer learning yields the best performance as the corpus size increases.

Change point detection: In data analysis, detection of change points is a tool to analyze the temporal patterns in a time series to detect the exact point of change in in future. From its inception back in the $50s$ [150] to recent days, with an abundance of its applications in various fields including speech processing, bioinformatics, climatology, finance and network traffic data analysis, a number of algorithms have been discovered in theory. Methods for change point detection (CPD) are roughly categorised into *online* [61] vs. *offline* [183], *univariate* vs. *multivariate* [114], and *model-based* vs. *nonparametric* [78, 174]. Online methods can be implemented in real-time setting in which algorithms run concurrently with the process being monitored and aim to detect the change point as soon as possible after it occurs. In the online setup, the algorithms need to inspect a batch of data samples (say ϵ data points which can be different in different methods) to be able to determine the change points between the old and the new data points. In

contrast, offline algorithms consider the entire time series at once and detect the change points in batch mode. Sometimes, the offline algorithms are called *signal segmentation* because segmentation is performed after the entire signal has been collected. In our work, we viewed the quality change of an article as a signal segmentation task and applied different offline methods to choose the best possible segmentation. In general, change point detection can be performed in either a parametric or a non-parametric framework. Parametric analysis necessarily assumes that the underlying distributions belong to some known family, e.g., Gaussian distribution etc. When the underlying assumptions of parametric models are largely unknown for the data at hand, non-parametric approaches can be deployed on any stream of continuous random variables without requiring any prior knowledge of their distribution. Popular methods include BinSeg [171], PELT [101] (in parametric setting) and ECP [134] (in non-parametric setting) for performing multiple change point analysis of multivariate observations.

2.1.2 Longitudinal analysis of Wikipedia

There exists several factors which directly influences the quality of articles, especially activity of the editors. The authors in [103] showed that effective coordination between editors leads to higher-quality articles on Wikipedia. Similarly, Liu and Ram [126] showed that coordination between editors and in particular, the specific roles of the editors ("who does what?") in the process of editing an article influences the quality of the articles. The community of editors [151] belonging to a specific wikiproject has a deeper sense of membership thus enhancing the durability of their contribution and finally leading to the overall growth of the Wikipedia as a whole. Some studies [12, 14] established the fact that the organizational structures in peer-production systems are not simple; rather different roles performed by the participants follow a career path which, in turn, confirms their stands in the community. [35] developed a model that could predict potential candidates to be promoted to the role of an *administrator*. A large number of works [89, 103, 104, 105, 186] reflected on the co-evolution of coordination and conflict in online knowledge production platforms and the factors behind growth of the conflict. They also illustrated coordi-

nation mechanisms to mitigate conflict and outlined the different aspects of successful decentralized environments. Further, collaborative measures [204] such as centralization, conflict and experience help to understand the crowd sourcing efforts in uplifting the quality of the articles. Moreover, the improvement is bi-directional; collaboration [107] helps in improving article quality, and in turn, it enriches the expertise of the editors. The authors in [100] through spatial and temporal analysis of edit history pointed out key duration of the day and week when anomalous edits are highly likely. In [13, 202] the authors showed distinct editor roles through the study of edit histories and demonstrated their impact on article quality. The authors in [98] showed that content retention and change in Wikipedia is similar to software development life cycles. Self-organisation of editors to create quality content generation in absence of explicit workflow constraint has been explored in [15, 97] through text and graph based approaches.

Researchers point "positive motivation" as an important pillar behind the success of the encyclopedia. Studies [11, 21, 70, 137] show different ways to measure the degree of labour-hours of editors and how their active participation determine the quality. In contrast, the conflict and bureaucracy of the participants can undermine the motivation of contributors [76], thus depleting the influx of contributors and ultimately declining the quality of the encyclopedia. To stop the decline of editors, researchers have suggested socialization policies and incentives [43, 84, 113, 138, 142] for the young as well as experienced editors in order to retain them. In this context, the ongoing Wiki education program [121] that includes students from different universities as new editors became successful enough and claims that institutionalized socialization works better in new editor retention. Besides the large pool of registered users, a significant number of anonymous users also participate in editing Wikipedia. The ongoing research [40, 67, 90] about the edit patterns of anonymous users in controlling the quality is a crucial direction. Although Wikipedia bots play an important role in reducing human labour, their activity cannot be ignored in diverse applications of automated patrolling tasks - quality monitoring, vandalism detection, etc. The studies in [71, 212] shed light on their activities, e.g., collaboration and conflicts among automated software agents.

Nowadays, OpenAI's GPT-4, Meta AI's Llama-3, and Google's Gemini have all been

released, marking a significant milestone in the competition among major tech companies to advance chatbot technology. The once unimaginable concept of using a chatbot to predict the quality of Wikipedia articles, explain its reasoning, and suggest improvements now seems much more achievable. While further research and development are necessary, and the future advancements of this technology remain unpredictable, it is clear that these tools have the potential to significantly enhance ongoing efforts in automatic quality assessment of Wikipedia articles.

2.1.3 Our work

There has been a large body of literature that has attempted to perform a deep dive into Wikipedia article quality monitoring systems. Our work adds the temporal dimension to the present literature on article quality; it attempts to characterize the quality life-cycle in the peer-reviewed system and also to predict the dynamic changes in the quality states early on. Using a subset of $30k$ articles, sampled randomly from English Wikipedia spanning more than 14 years from the creation of the articles, we build a data pipeline to observe the temporal evolution of the article quality. We perform rigorous experiments to understand the fine-grained details of the quality change phenomenon, which we elaborate on along with their implications in Chapter 3. The closest work to this is by Zhang et al. [206], where they identified three types of quality trajectory, i.e., stalled, plateaued, and sustained over a dataset of 6000 articles. One of the major drawbacks of their approach is that the authors did not use the actual quality of the articles by leveraging the Wikipedia dumps, which are evaluated by human experts. Instead, they used ORES [75] to predict the quality labels, which can be potentially erroneous because it does not take into account the temporal changes of article attributes. Moreover, with the static frameworks of calculating quality at a particular timestamp, these models achieve at best 63.5% accuracy [50, 73]. We, on the other hand, take a relatively simpler approach of *multivariate change point detection* [184] where we aim to detect a quality change as a function of a set of very intuitive features. This approach is significantly lightweight compared to explicit quality prediction approaches [73, 176], which train models with millions of parameters.

Moreover, our prediction model outperforms the existing machine-learning-based quality prediction framework of Wikimedia Foundation, ORES by a significant margin.

2.2 Knowledge equity across multiple languages

Wikipedia has been studied and modeled by researchers for over 20 years, and it is not a monolithic resource. By 2005, there were already almost 200 language editions of Wikipedia and, as of 2023, there are over 300 Wikipedia versions exist[1]. Over the past few years, the research community has expanded their work to be more multilingual – including more languages, especially non-English and smaller language editions of Wikipedia.

2.2.1 Multilingual Wikipedia research

Many researchers have examined differences between different language editions of Wikipedia from the standpoints of content (i.e., text, image), readers [18], and editors [27] as well. Text diversity in Wikipedia has collectively demonstrated that textual content about the same concept is highly diverse across language editions. This leads to considerable information gaps, creating an information divide between users of different languages Wikipedias [62]. The English Wikipedia is the largest edition, with over 5.8 million articles. In contrast, the information coverage in other language editions, even for widely spoken languages, is only a fraction of the content available in English Wikipedia.
Information asymmetry in multilingual Wikipedia: Authors in their work [81] showed language's fragmenting effect on user-generated content by examining the diversity of knowledge representations across 25 different Wikipedia language editions. One way to compare the different editions of Wikipedia is in terms of the coverage of concepts (or entities). Often, it is assumed that a single Wikipedia article corresponds to a unique concept (or topic or entity), and the overlap of these concepts is a useful measure to compare different Wikipedia versions. For example, authors [62] considered a set of 48

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

people in the DUC 2004 biography generation task and studied in how many languages these people have Wikipedia entries and compared their length. Different language editions of Wikipedia serve very different communities [96, 115] and thus often cover very different topics [23, 136]. The resulting variation in quality and quantity of content about different topics [119] presumably would affect the resulting vocabulary, hence reducing the ability of multilingual language models trained on Wikipedia to handle different topics accurately. Due to the diverse cultural backgrounds of Wikipedia editors, there is a bias in content selection based on local culture and more in-depth knowledge of local facts [37, 80]. This contributes to variations in the content across different language editions of Wikipedia. For example, authors in [79] showed that although language editions cover the same topics, editors often choose culturally relevant imagery to include into it. Previous studies have revealed the behavior of editors and found that only about 15% of of them edit multiple editions of Wikipedia [74]. Further, it is observed that editors contribute more complex information in their native language [154]. Being the center point of diverse cultures, Wikipedia has gained unprecedented success by leveraging its distinctive ideologies. However, this individual ideological stance of editors leads to cross-lingual violations of the NPOV. Authors in their work [72] found that specific lexical and visual choices by editors are ideologically motivated and go against the principles advocated by NPOV. Researchers in [166] quantify and provide valuable insights into the information gap between different language editions of Wikipedia. They offer a roadmap for the information retrieval community to help bridge this gap.

2.2.2 Cross-lingual knowledge transfer

Wikipedia data has been utilized as part of the training datasets for popular (multilingual) language models like multilingual BERT [157]. Such a training setup helps to widen the scope and improve the accuracy and inclusiveness of multilingual models and their applications. For this purpose, content alignment and content transfer among different language versions, especially with low-resource seem extremely crucial. A large body of work focused on addressing the information asymmetry between different language

editions of Wikipedia. [19] et. al. proposed the RECOIN system to measure the completeness of information about an entity by using other similar entities as a reference. The work by [199] described an algorithm that identifies articles missing in a target language based on a source language. These missing articles are ranked by their expected future page views and recommended to editors according to their interests. [23] presented a system that aggregates information about a concept from multiple language editions of Wikipedia for end-users. Authors in [2] developed an automated system to align infoboxes about an entity across different Wikipedia versions. This system can create new infoboxes or fill in missing information in existing ones using data from infoboxes about the same entity in other languages. In another study [32], authors presented a method for cross-lingual alignment of template and infobox attributes in Wikipedia. The alignment is used to add and complete templates and infoboxes in one language with information derived from Wikipedia in another language. In the case of languages with limited or poor translation resources, authors [153] proposed a lightweight approach to measure cross-lingual similarity in Wikipedia using section headings rather than the entire Wikipedia article, and language resources derived from Wikipedia and Wiktionary to perform translation.

Generative approaches: Existing methods like [169] use the high-level structure of human-authored texts to automatically generate domain-specific templates for obtaining new overviews. Authors in [161] used Wikipedia articles in nine languages to identify word translations through the use of keywords and a word alignment algorithm. With the advancement of generative AI, recent works have focused on creating new Wikipedia pages in low-resource languages from scratch. Authors in [4] introduced a novel framework for creating Hindi Wikipedia pages in the 'Scientific Person' domain for scientists who do not yet have a Hindi Wikipedia page. This framework generates template sentences from information collected via Wikidata to fill in information for any scientist, and the authors argued that their framework significantly outperforms Wikipedia's internal translation system. Another work [181] focused on cross-lingual multi-document summarization of text from multiple reference articles in various languages to generate Wikipedia-style text. The authors contributed a benchmark dataset, XWikiRef, which includes around 69K Wikipedia articles covering five domains and eight languages. Using this dataset, they trained a two-stage system where the input is a

set of citations and a section title, and the output is a section-specific summary generated in low-resourced languages like Hindi.

2.2.3 Our work

To address content disparities across different language editions of Wikipedia, we introduce a lightweight framework aimed at enhancing knowledge integrity across diverse linguistic communities in Chapter 4. We have considered English as high-resourced and Hindi as low-resourced language in our study. Current state-of-the-art studies utilize generative AI systems to reduce the information gap across multiple language versions of Wikipedia. However, this approach can introduce misinformation [214] and social biases [69, 110] from large language models, violating Wikipedia's NPOV policy. Instead of directly employing generative AI systems, our framework extracts relevant content from external resources available in high-resource languages, such as English. This extracted content is then adapted to match Wikipedia's distinctive style, particularly its NPOV policy, using the in-context learning capabilities of large language models. In addition, we incorporate extract knowledge from external English resources into our framework. Finally, the combined knowledge (both existing and external content) in the high-resource language is machine-translated into the target low-resource language, making it ready for further integration into Wikipedia articles in that language. Our lightweight framework is efficient in generating new content for individual Wikipedia sections. Further, our rigorous evaluation methods, both based on heuristics and human assessment achieves a significant improvement in adapting new content in individual sections of low-resource languages.

2.3 Societal biases in Wikidata

2.3.1 Knowledge graph embedding

KG embedding algorithms have been developed to learn compact representations of entities and relations within a KG in a low-dimensional embedding space. These embeddings aim to capture the semantic relationships between entities and relations by training an objective function that maps a triple to a scalar score, maximizing the likelihood of correct prediction of the triples. Representation learning on KGs is a highly active direction in research, with numerous novel KG embedding (KGE) algorithms being proposed recently, including *TransE* [30], *TransD* [92], *TransH* [194], *RESCAL* [144], *DistMult* [200], *HolE* [145], *CrossE* [209], *ComplEx* [182] etc. Simultaneously with the recent progress in neural network approaches, several methods have been proposed using convolution kernels, among which *ConvKB* [143], *ConvE* [56], *HypER* [20] are the important ones. On the other hand, various methods [46, 112, 210] have been proposed to perform link prediction, relying on different embedding representation techniques [165, 188].
TRANSE [30], a foundational model for learning KG embeddings, employs a geometric approach implementing a translation operation to generate the embedding of a tail entity based on the embeddings of the head entity and the relation. Given a triple $< h, r, t >$, TRANSE aims to learn vectors h, r, and t in such a way that the distance between $h + r$ and t is minimized for positive edges. Conversely, for negative edges, TRANSE strives to learn embeddings that keep $h + r$ far from t. $f_{TransE} = -||h + r - t||_n$.
DISTMULT [201] assumes the relation r as a diagonal matrix and follows the scoring following function – $f_{DistMult} = h^T diag(r)t$. This score captures pairwise interactions between the components of h and t, thus allowing the model to emphasize or de-emphasize parts of the vector representation based on the relation.
For COMPGCN [185] method that combines state-of-the-art knowledge graph embedding learning algorithms, for example, TRANSE, DISTMULT, etc., with popular graph convolution network (GCN) for incorporating multi-relational information in knowledge graphs. Further, COMPGCN can scale with a growing number of relations in a relational graph and can be trained on multiple downstream tasks. GEKC [127] is the latest addition to the

KGE learning paradigm, which uses generative KGE circuits to enhance the efficiency and reliability of triple predictions for the missing link prediction task. Our study aims to demonstrate how models from each of these genres perform on our curated dataset.

2.3.2 Existence of social biases

As often KGs are projected as collaborative repositories, it is quite understandable that they would be subjected to human perception and cognitive biases. Such social biases are reciprocated in terms of the distribution of entities and relations and get embedded in KGEs. Recent work [55] on finding data biases in collaboratively constructed knowledge graphs, especially in Wikidata investigates how paid crowd-sourcing can be used to understand contributors' implicit bias. Specifically, the authors in this paper recruited crowdworkers to verify controversial facts and demonstrated the benefits of surfacing bias information to end users of applications rather than eliminating them from the knowledge graph. Authors in [205] find under representation of content about women as compared to men in Wikidata. Following the general predominance of the male population around the globe, Wikidata editors add many male-dominated occupations. However, Wikidata is no more biased than the real world; it mirrors the existing gender gap in our society. Similar work [172] investigates the presence of race and citizenship bias in Wikidata. The study reveals that White individuals and those with citizenship in Europe and North America are over represented in comparison to the rest of the world. In similar lines, there exists research works [203], attributed to Wikipedia, a sister project of Wikidata in which authors tried to reveal how imbalanced the gender presentation is on the occupations pages of Wikipedia. Authors in [83] studied gender differences in various Wikipedia (European) language editions with respect to the coverage of the Members of the European Parliament (MEP). For further investigation, they inspected differences in the content of Wikidata entries of male and female MEPs and found gender imbalance across nationality. Besides the representation biases, researchers intended to characterize and mitigate the inference biases arising out of sensitive attributes, such as gender, ethnicity, religion etc. to finally make the KGE learning algorithms bias-free.

Several works [33, 65] have successfully conducted experiments to show that harmful biases are penetrating societal spheres. Authors in a recent work [99] have suspected human assumptions related to the choice of sensitive relations. As a result, they proposed a framework that is capable of identifying biased attributes automatically based on some metrics. In an orthogonal direction, researchers are trying to invent various methodologies to mitigate biases from knowledge graphs which in turn will be helpful in designing a bias-free automated system for different machine learning and NLP tasks. As a useful solution, Arduini et al. [17] developed a debiasing method based on adversarial learning that modifies the embedding by filtering out sensitive information but preserving all the other relevant information. In [64] the authors presented a novel approach in which biased embeddings are trained to be neutral with respect to some sensitive attributes, such as gender-based on adversarial loss, and later users are allowed to add sensitive information back to the system on demand. This method has been proven to be significantly faster and more accurate than previous approaches [31] of debiasing knowledge graph embedding.

2.3.3 Fairness in link prediction

Despite the importance of link prediction, biases can occur due to various factors such as incomplete data, imbalanced training data, or the presence of implicit biases in the training data or the link prediction algorithm itself [193]. Such biases can lead to inaccurate or unfair predictions, particularly in applications such as recommendation systems or decision-making algorithms [42]. Social biases, such as gender bias, can lead to various challenging problems, particularly in large peer-production platforms like Wikipedia [36, 187] and Wikidata. In [159], the authors identified class-level knowledge gaps in Wikidata. Most of the above studies have been US-centric. We know only one work that considered race and country of citizenship bias in Wikidata [172]. In the context of link prediction, fairness denotes that the predictions should not systematically discriminate against particular individuals or groups based on certain sensitive attributes such as gender, race, religion, age, etc. Several link prediction models [38, 124] have been proposed to ensure fair link prediction in social network graphs.

2.3.4 Our work

The studies so far discussed [33, 63] have depicted biases that persist in KGs, as well as employed several metrics for measuring the biases. Further, many studies [17, 64] have proposed debiasing strategies to remove the social biases from the current version of the KGs. However, such studies lack the systematic exploration of the sensitivity of the bias measurements through varying sources of data or the embedding algorithms used. To address this research gap, we present a holistic analysis of bias measurement on the knowledge graph in the first work (section 5.1) of Chapter 5. We analyze the knowledge graph obtained from thirteen different geographies spanning seven continents and identify the biases that surface in Wikidata. Next, we unfold the variance in the detection of biases using two different knowledge graph embedding learning algorithms - TRANSE and COMPLEX. We conduct our extensive experiments on a large number of professions sampled from the thirteen geographies with respect to the sensitive attribute, gender. Our results show that the inherent data bias that persists in KG can be revised by specific algorithm bias as incorporated by KG embedding learning algorithms. Further, we show that the choice of the state-of-the-art KG embedding algorithm has a strong impact on the ranking of biased professions irrespective of gender. In particular, we find that the embedding algorithm COMPLEX is more robust to the choice of geographies compared to TRANSE. Subsequently, we observe that the similarity of the biased professions across geographies is minimal, which possibly reflects the socio-cultural differences around the globe. This kind of observation is often overlooked by most of the coarse-grained approaches working at the aggregate level.

In the second work (section 5.2) of Chapter 5, we address bias and fairness issues in the link prediction task, which is crucial for automatic knowledge graph completion. To understand this problem in detail, we curate a large dataset of 3.2 million human entities with over 30 thousand different professions sourced from the latest (December 2022) Wikidata dump. Our meticulously curated dataset comprises human entities from 21 different geographies around the globe. Subsequently, we introduce a framework – AUDITLP – to identify biased outcomes in link prediction, specifically how occupations

are classified as either male or female-dominated based on gender as a sensitive attribute. Similarly, we experiment with the sensitive attribute of age and observe that occupations are categorized as young-biased, old-biased, and age-neutral. Our benchmarking experiment reveals nuanced micro-level characteristics of gender/age-biased occupations for different geographies in our dataset. Further, we show that the variance in the biased outcomes across geographies neatly mirrors the socioeconomic and cultural division of the world, resulting in a clear partition of the Global North from the Global South.

Chapter 3

Assessment of quality changes in Wikipedia articles

In this chapter, we conduct a systematic analysis of the quality evolution of English Wikipedia articles. In addition, we present a novel unsupervised approach that utilizes change point detection algorithms to predict the upcoming changes in the quality of Wikipedia articles.

3.1 Quality evolution of Wikipedia articles and temporal patterns

Knowledge accumulation and dissemination through peer production as a form of collective intelligence, working in a decentralized manner without the inclusion of any relational or contractual workforce, is a predominant source of social good, emerging out of the *world wide web*. With the speedy growth of the voluminous amount of their content, ensuring the quality and reliability of the content processed through the collective system [196] is a challenging task. The quality can be assessed in various scales

and standards that help in earning the trust of the platform among the consumers. Hence, over the days, researchers are engineering different aspects of quality, which in turn is responsible for the increase in the viewership of the peer-production system. Likewise, the platform's success is directly associated with quality, and quality, in turn, is positively correlated with collaboration patterns within the system. The crowds tend to change their contribution with varying incoming and outgoing participation, and hence, the quality itself is dynamic and unstable over time. Understanding the temporal evaluation of the quality of large-scale peer-production systems, such as Wikipedia, avail researchers with important clues about the system's success.

Every day, millions of participants, comprising both creators and consumers, pour into Wikipedia to keep themselves updated on a plethora of topics. The platform is benefited from this synergy and is grown in terms of volume and veracity over the last decades. In this work, we focus on the English Wikipedia, which is the largest wikiproject covering more than $\sim 6M$ articles and an astonishing $\sim 3B$ words[1]. Wikipedia's bold policy, "anyone can edit," draws the attention of the crowd, irrespective of the socio-cultural and geographical boundaries, to collaborate and contribute openly. Many of them, especially the editors who are concerned with the quality standards, assign the articles to the existing quality scales, but the quality keeps on fluctuating in the strongly moderated review environment by individuals or panels. Wikipedia's openness often leads to information manipulation and vandalism by inexperienced editors and vandals. In addition, with the vast volume of articles in Wikipedia, most of them are not updated periodically, hence becoming inconsistent and incomplete. Both are serious concerns leading to the potential degradation of the encyclopedia quality. To overcome this shortcoming, Wikipedia has implemented a user-driven approach to assess the quality of articles. According to Wikipedia's guidelines, an article can be evaluated to any of the quality rankings (FA, A, GA, B, C, Start, Stub; ordered in terms of decreasing quality) by the community of editors. Although the manual assessment includes the perfection of quality assignment tasks, a major problem that often comes up is of inconsistency. For example, about $4,54,697$ articles have remained unassessed till 2020 in the English Wikipedia. Moreover, the

[1] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

assessment becomes obsolete quickly with the frequent updation of information. Hence, in reality, the quality of an article is not a static attribute; in contrast, it follows a temporal trajectory of improvements and declines. In the last few years, the Wikimedia Foundation has started searching for automatic solutions for quality evaluation designed for Wikipedia specifically. Leveraging diverse collaborative features of this peer-review system and the structural features of the content, several AI-based (state-of-the-art machine learning and deep learning approaches) techniques have been developed to measure the quality of an article at a specific timestamp. The later solution is able to eradicate the time lag of manual effort but introduces noise in measuring the dynamic change. Depending upon the cost of the system, the automatic prediction framework enables the models to be learned on the entire dataset or partially on a few sets of articles. Hence, they fail to capture the dynamic changes of every article individually and create prediction errors. In this work, we try to fill the gap between the two approaches - manual and automated- by incorporating dynamic and unstable changes in quality. Our work adds the temporal dimension to the present literature [50, 85, 198] on the article quality and tries to characterize the quality life-cycle in the peer-reviewed system and also to predict the dynamic changes in the quality states from beforehand. Using a subset of the $30k$ articles, sampled randomly from the English Wikipedia articles spanning more than 14 years from the creation of the articles, we build a data pipeline to observe the temporal evolution of the article quality.
Research questions: We are motivated with the following research questions–

1. **R1:** How do the Wikipedia articles transition through different quality states over time?

2. **R2:** What will be the appropriate way to detect the dynamic change in article quality? What are some of the most intuitive characteristics of an article that help in forecasting the upcoming changes in its quality?

3.1.1 Our contribution

Quality change evolution: We find various interesting temporal patterns in the quality of the articles. There is a group of articles that undergo continuous improvement in quality, finally reaching the highest quality category: featured articles or good articles. Another group of articles undergoes interspersed stages of quality improvement, sometimes even making sudden jumps from a very low-quality category to a very high-quality category and vice versa. In many cases, we observe that there are cycles formed, i.e., starting from a quality category the page undergoes a series of quality changes and comes back to the same category that it started from (aka cyclic switches). Surprisingly, almost half of the articles in our dataset remain in the same quality class all through their lifetime without undergoing any promotion or demotion. Almost all of these stagnant cases are low-quality articles.

Change point detection: Apart from designing rigorous experiments to understand the phenomenon of quality life cycle in Wikipedia, we leverage state-of-the-art multivariate change point detection algorithms [184] to solve a novel task of quality change point detection. We combine a series of features from editor attributes, article attributes, and activity-based features, which helps us achieve 76% *covering* in the detection of quality change points. We observe that at an aggregate level, the content of an article digested through a set of article attributes is the prime determinant of its quality. Some of these include the length of the article, the number of references in the article, the number of images, the number of links to other articles, the presence of an infobox, etc. However, when we deep dive, we observe that a mix of organizational attributes of the article, like the number of revisions on the talk page, the mean time elapsed between two revisions on the talk page plus a special content attribute of the article, i.e., the ease of its readability in terms of the number of difficult words act as an even better predictor of quality change points. The effectiveness of such nuanced feature combinations is hitherto unknown. Our approach is unsupervised and page level, unlike traditional approaches of quality prediction where features across pages are used as inputs to a machine learning model [175] for predicting quality. Furthermore, our prediction model makes use of these attributes to far

outperform the state-of-the-art models of end-to-end quality prediction like ORES by a significant margin. We aim to use our change point detection approach to keep potential editors updated about the possibility of quality change and generate suitable alerts for them whenever appropriate.

3.2 Dataset

3.2.1 Wikipedia article quality assessment

Several critics of the peer-production system question the quality of the content - whether the large-scale collaboration is potentially capable of maintaining quality standards. Despite the variety of challenges, the peer-production systems try to address the quality issues periodically through manual or automated reviews, and Wikipedia follows the same path. Wikipedia articles exhibit a wide range of quality, from comprehensive, fully-referenced, and well-illustrated articles that thoroughly cover their topics to brief stubs consisting of just a single sentence that merely defines the subject. Accurately distinguishing between these extremes and the intermediate quality levels is highly beneficial. To achieve this, Wikipedia editors have established detailed rubrics[2] based on criteria such as topic coverage, organization, and technical style for assessing the quality of articles. These quality levels guide editors in evaluating and prioritizing their efforts, while researchers use these assessments to analyze content dynamics. Furthermore, developers use these quality levels as filters when building recommender systems or other tools. For example, in English Wikipedia, the different quality classes are FA[3] (featured article), A[4], GA[5] (good article), B[6], C[7], Start, and Stub. Here FA is placed at the highest rank - articles that are fairly complete and well written. Stub, on the other hand,

[2]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment
[3]https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria
[4]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment/A-Class_criteria
[5]https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria
[6]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment/B-Class_criteria
[7]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment/C-Class_criteria

has the lowest in quality - very little meaningful content with a need for improvement in the article content. In other words, a very short article, containing at least two meaningful sentences about the topic is usually designated as Stub, and slightly longer articles are placed in the Start class. The grading scheme intends to verify a few criteria as follows–

- Type of the content: articles should be well-written, comprehensive, well-researched, neutral, and stable.
- Organization of the content: articles should have a lead section briefing the topic, hierarchical structure of sections, sub-sections, and consistent citations.
- Inclusion of images with appropriate captions and acceptable copyright status.
- Articles should maintain a decent length without unnecessary details.

Similar quality divisions exist in other language editions, like in the French Wikipedia, where quality classes such as AdQ, BA, A, B, BD, and ebauche are used, in which AdQ represents the highest-quality article, and ebauche stands for lowest quality, similar to Stub in English Wikipedia. The corresponding quality class is mentioned on the talk page of an article. Talk pages act as discussion platforms where editors collaborate and exchange ideas about the quality of articles. Typically, assessments on lower-quality classes are mainly controlled by the organized group of editors, such as Wikiproject communities[8], who tag the quality changes on the talk pages. The quality ratings documented on the talk pages are subsequently gathered and logged as statistics by automated bots. For achieving the highest quality (FA or GA), potential articles are nominated by the editors and later reviewed by the team - individuals or panels. Articles meeting the criteria for these distinctions are selectively added to the Featured Articles (*WP:Featured articles*) or Good Articles (*WP:Good articles*) lists. Wikipedia maintains individual lists of articles that have satisfied all the criteria of *featured*[9] or *good*[10] articles in the review

[8]https://en.wikipedia.org/wiki/Wikipedia:WikiProject
[9]https://en.wikipedia.org/wiki/Wikipedia:Featured_articles
[10]https://en.wikipedia.org/wiki/Wikipedia:Good_articles

process. The community periodically updates these lists to ensure their accuracy and relevance. Such maintenance works need to be performed in a timely manner to satisfy the trust issue of viewers. However, auto-patrolling is not normative in Wikipedia so far, and therefore, a large volume of articles misses the attention of the peer-review community.

3.2.2 Dataset Description

Wikimedia foundation provides access to the complete revision history of all the articles of different language versions in the form of Wikidumps[11]. For our work, we download the first 100 English dumps which are stored as 7z archived xml files. These dumps consume $\sim 8TB$ disk space in uncompressed form and consist of $\sim 6m$ English Wikipedia pages. Each uncompressed xml file of size $\sim 80GB$ contains a random collection of $\sim 5k$ Wikipedia pages. Because of its periodic updation, the pages have all the revisions from the date of creation to the last version as of June 2019. We parse each xml file by the mediawiki xml[12] parser to find out a sample of articles of *main*[13] namespace as well as the corresponding talk pages in one linear scan. Specifically, let us assume, the parser encounters a main article P; we remember it and try to locate *Talk:*P in the later scans or vice-versa. However, if the talk page of P is not present in the current xml file or the other downloaded dumps, we ignore the page P in that case, otherwise we include the page in our dataset. While continuing this process, we keep track of the number of articles in each of the quality classes: **FA**, **GA**, **A**, **B**, **C**, **Start** and **Stub** so as to maintain the class proportions. Although the above mentioned process is quite simple, it is able to extract an approximately balanced number of articles from each of the quality classes (see Table 3.1). As the **FA** and **A** articles are limited in number, we do not follow the omission approach outlined above for these two classes; instead we include all of them.

The articles we collect through the above mentioned process are free from any kind of selection bias except maintaining the ratio of the articles in the individual quality classes.

[11] https://dumps.wikimedia.org/

[12] https://pypi.org/project/mwxml/

[13] This contains all the encyclopedia articles, the lists, the encyclopedia redirects etc.

Furthermore, to verify the generalizability, we compute the category distribution of our $30k$ articles and compare the same with the actual one of the entire English Wikipedia. The articles in our dataset consist of 102666 categories among the existing $1m$ categories of Wikipedia. Further, we rank the frequently occurring categories, in which *living people* and *American films* are the top ones, covering 20% articles in our dataset. If we consider the whole set of English Wikipedia articles these two categories are again at the top making 16% of the whole data. We compute the Spearman's rank correlation coefficient (ρ) for the two rankings (truncated at top five categories constituting respectively 24% and 17% of our dataset and the full Wikipedia) and obtained a ρ value as 0.9. Thus, we show that our dataset and the results we report based on it should be representative sample of the English Wikipedia.

Class							**Total**
FA	A	GA	B	C	Start	Stub	
3536	511	5780	5335	4884	5459	5321	30826

Table 3.1: Count of articles in the respective quality classes.

3.2.3 Merging of quality classes

First, the content of Wikipedia articles and the talk pages in the form of Wiki Markup Language[14] (wiki text) are converted into plain English-like text using a standard python text crawler[15]. Special tokens are then used to identify meta contents in the page such as *infobox*, *level 1 section headings*, *level 2 section headings*, *interal wikilink*, *external link*, *inline references*, *footnote template*, *quotation template* and the categories using the mediawiki parser[16]. According to the hierarchy[17] of quality class division provided by Wikipedia, every class has a related detailed criteria of standards; however, the classification is based on qualitative measures primarily. For example, in the description of the

[14]https://en.wikipedia.org/wiki/Wikipedia:Wiki_Markup_Language
[15]https://github.com/attardi/wikiextractor
[16]https://mwparserfromhell.readthedocs.io/en/latest/
[17]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

C class, an instruction in the guidelines reads "*The article should have some references to reliable sources, but may still have significant problems or require substantial cleanup* . . . " with some indefinite quantifiers. So, it is difficult for machine learning models to distinguish the quality classes from the immediate upper/lower classes. Further, there is extreme sparsity of data in some classes which might not allow us to obtain statistically reliable insights. We therefore club the quality classes as noted in Table 3.2. According to our clubbing scheme the four types of quality classes we arrive at are **FA**, **AGA**, **BC**, **SS** which can be arranged in increasing order of qualities as follows

$$\mathbf{FA} > \mathbf{AGA} > \mathbf{BC} > \mathbf{SS} \tag{3.1}$$

Note that this new ordering does not violate the predefined hierarchy of quality classes as determined by the Wikipedia community.

Old class	New class	Counts
FA	**FA**	3536
A,GA	**AGA**	6291
B,C	**BC**	10219
Start,Stub	**SS**	10780

Table 3.2: Count of articles of the newly defined quality classes after merging.

3.2.4 Basic characteristics of the quality classes

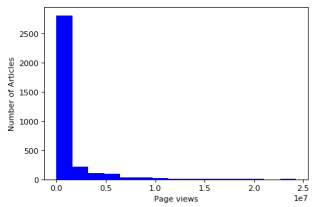
In this section we investigate the characteristics of the high-quality and low-quality articles in terms of (i) topics, (ii) user views, (iii) editions/revisions, (iv) number of collaborators as suggested by the reviewer. Here, we denote the articles belonging to FA and AGA quality classes as the high-quality and the articles of SS class as the low-quality ones.

Topics: We rank the frequently occurring categories in which *living people* and *Americans films* are the top ones. We try to find similar rank in high and low quality articles, in which high quality articles attain the same ranking as that of the whole dataset.

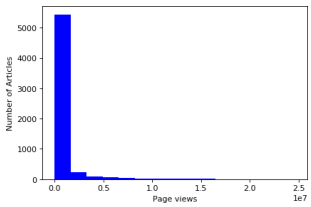
User views: We collect the page views of every article from the Wikimedia API[18] from the date of availability till June 2019. As expected, the articles belonging to high-quality classes draw more user views than the lower ones. The distribution of user views of every class are plotted in the Figure 3.1.

Editions/revisions: We try to compare the articles in terms of number of editions/revisions that an article has passed through from the date of the creation. In Figure 3.2, we plot the distribution of editions for the high and low quality articles in which high quality articles (FA, AGA) show larger number of revisions in their life-cycle.

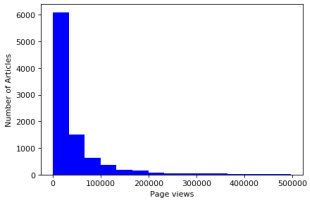
Number of collaborators: Similar to the descriptors as mentioned earlier, we observe the distribution of the number of collaborators for high and low quality articles in Figure 3.3. The high quality articles employ larger number of collaborators as compared to the lower quality ones.



(a) User views: FA class.

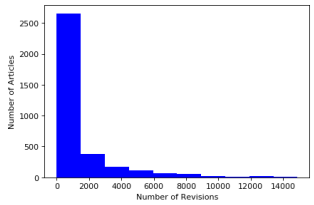


(b) User views: AGA class.

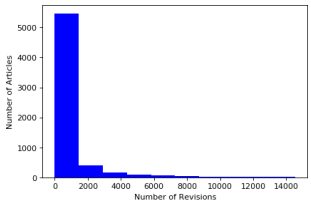


(c) User views: SS class.

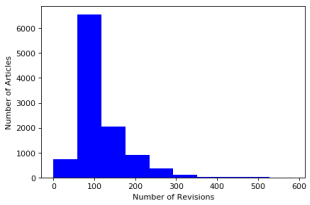
Figure 3.1: Histogram showing user views of articles belonging to high (FA, AGA) and low (SS) quality classes.



(a) Editions: FA class.



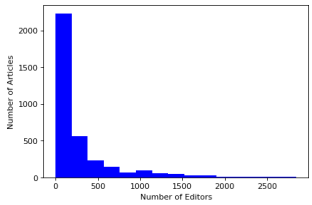
(b) Editions: AGA class.



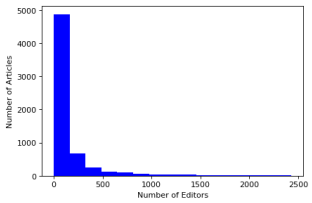
(c) Editions: SS class.

Figure 3.2: Histogram showing number of editions/revisions of articles belonging to high (FA, AGA) and low (SS) quality classes.

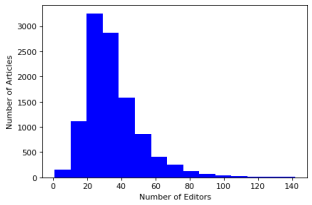
[18]https://wikimedia.org/api/rest_v1/#/



(a) Collaborators: FA class.



(b) Collaborators: AGA class.



(c) Collaborators: SS class.

Figure 3.3: Histogram showing number of collaborators, i.e., editors of articles belonging to high (FA, AGA) and low (SS) quality classes.

3.3 Temporal evolution of article quality

Although measuring Wikipedia article quality is a complex venture, the quality assessment task by the editorial team of typical articles or Wikiprojects prioritizes further contributions to the articles, influencing editors to revamp continuously. In this section, we carry out a detailed analysis of the typical temporal patterns exhibited by the change of quality classes of the articles in our dataset. We adhere to the quality classes restructured by us, as outlined in the previous section (refer to the subsections under Section 3.2). While articles within a given quality class may vary slightly in terms of some specific quality factors, we assume that all articles within the same quality class are comparable in terms of overall quality. Based on this assumption, we analyze the temporal patterns demonstrated by changes in the quality classes.

3.3.1 Only promotion

Let us assume, an article P had the quality **SS** assigned to it at timestamp t_x, the quality **BC** at timestamp t_y and finally if the current quality of the article P is **FA**, where $t_x < t_y$, we categorize the change of quality as constant *promotion* and the article P is placed in this category. Note that promotions can also skip intermediate classes (e.g., direct **SS**⇒**FA**). Following this criteria, we observe that 13249 articles experience only promotions in qualities over the revisions and the detailed statistics is noted in Table 3.3.

A page with the sequence of temporal change **SS** ⇒ **BC** ⇒ **FA** is counted twice in the sub-categories **SS** ⇒ **BC** and **BC** ⇒ **FA** individually. We observe that 42.97% articles in our dataset undergo one or more promotions. Surprisingly, we found a number of articles falling under the sub-category **SS** ⇒ **FA** have the mean time (avg. time) for quality change more than 5 years. The average time taken for promotion to **FA** status does not appear to be significantly influenced by the article's topic. A number of articles, for example, *Pinkerton (album)*, *Coenwulf of Mercia*, *Denbies*, *Duncan Edwards*, *Nativity (Christus)*, *Roy Welensky*, *Aspasia*, etc. achieved promotion from **SS** to **FA** in approximately one year, which is notably faster than the reported average time frame. While our dataset includes articles mostly from categories such as *living people*, *American films*, no clear topical influence on promotion time is observed. However, it is reported that paid editing can sometimes accelerate quality assessment improvements[19], aiding articles in achieving FA status and subsequently being featured as *Today's Featured Article*[20]. Although the standard deviation (SD) reported in this sub-category is slightly higher compared to some other sub-categories (e.g., **SS** ⇒ **BC**, **SS** ⇒ **AGA**), it is quite evident that many pages go unnoticed for as long as 5 years or even more resulting in very delayed quality assessment. The promotion from **SS** to **BC** takes the second highest average time. The fastest category upgradation (in terms of mean and standard deviation of average turn around time) is naturally from **AGA** ⇒ **FA** which attracts the largest attention of the editors.

Type	Number of hops	Count	Avg time (in days)	SD (in days)
SS ⇒ **BC**	1	8594	1253.70	1165.91
BC ⇒ **AGA**	1	6144	390.58	728.73
AGA ⇒ **FA**	1	2283	294.18	476.34
SS ⇒ **AGA**	2	1384	1198.25	1206.46
BC ⇒ **FA**	2	487	535.28	881.65
SS ⇒ **FA**	3	97	1873.44	1380.87

Table 3.3: Count of articles with only promotion. Highlighted rows show changes that draw special attention.

[19]https://tinyurl.com/5bvpvy7a
[20]https://tinyurl.com/3f3hy8dc

3.3.2 Only demotion

We observe that the temporal sequences of 221 articles in our dataset undergo one or more demotions in quality classes. Likewise *only promotions*, a fall in any number of quality classes is considered as *only demotion*. Table 3.4 shows overall statistics of this category of temporal change. As in the *only promotion* category, a page with demotions greater than one hop is counted in each of the sub-categories individually.

Although only a handful of articles get demoted over the time, such demotions specially for the sub-category **FA** $\Rightarrow$ **BC** with an average time gap of quality assessment more than 2 years indicates a surprising exception in article quality changes in Wikipedia. On manual inspection we find that the **FA** pages in this category lack quality contribution for a long time and are hence demoted to lower classes.

Type	Number of hops	Count	Avg time (in days)	SD (in days)
BC $\Rightarrow$ **SS**	1	83	542.86	655.51
AGA $\Rightarrow$ **BC**	1	105	400.33	401.32
FA $\Rightarrow$ **AGA**	1	2	469.88	294.94
AGA $\Rightarrow$ **SS**	2	1	21.92	0
FA $\Rightarrow$ **BC**	2	32	753.54	694.44
FA $\Rightarrow$ **SS**	3	0	0	0

Table 3.4: Count of articles with only demotion. The highlighted rows denote rare cases of demotion in quality.

3.3.3 Both promotion and demotion

This set of pages indicates a sequence of promotion and demotion over the revisions in their lifetime. There are 1407 such articles in our dataset.

3.3.4 No change in quality

Surprisingly, there exists a large fraction of articles that have undergone the quality evaluation process only once in their entire lifetime. As a result, the quality class assigned to these articles by the evaluators remains unchanged throughout the analysis period. While some quality improvements might have occurred through edits, the assessed quality class remains static over time due to the lack of subsequent evaluations or ground truth assessments. Therefore, within the scope of our analysis, we consider the quality of such articles as constant and classify them in the group with no quality changes. We include such 15949 pages in this category, and the majority of these pages come in the quality class **SS**. The distribution of articles of different quality classes is mentioned in Table 3.5. The mean time denotes the average time gap between the creation time of the pages and the time of the first quality assessment. **FA** pages experienced the highest mean time for the first evaluation. The existence of a large fraction (51.73%) of the articles in this category clearly depicts that the majority of the articles are overlooked by the current quality assessment framework, which is possibly becoming a norm on this platform. Overall this may not be a good sign for the health of the platform.

Quality Class	Count	Mean time (in days)	SD (in days)
FA	341	2039.14	1750.22
AGA	375	1165.19	1610.10
BC	4586	935.45	1075.24
SS	10647	316.02	581.87

Table 3.5: Count of articles with no change in quality.

The **alluvial diagrams** in Figure 3.4 and 3.5 present an illustration of the temporal change of quality classes in two different time windows: 2010-2014 and 2014-2019. The figures further visually make it evident that more often than not the quality of an article remains fixed over time; only in rare exceptions they exhibit a change.

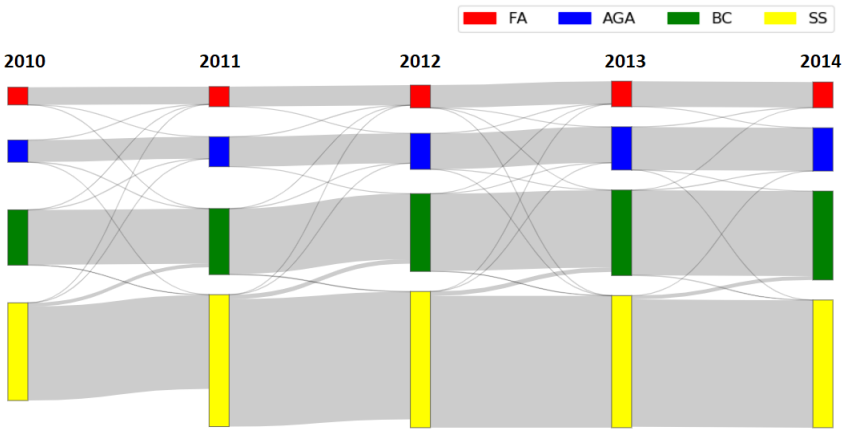


Figure 3.4: Temporal illustration of intra and inter class quality changes over the years 2010-2014.

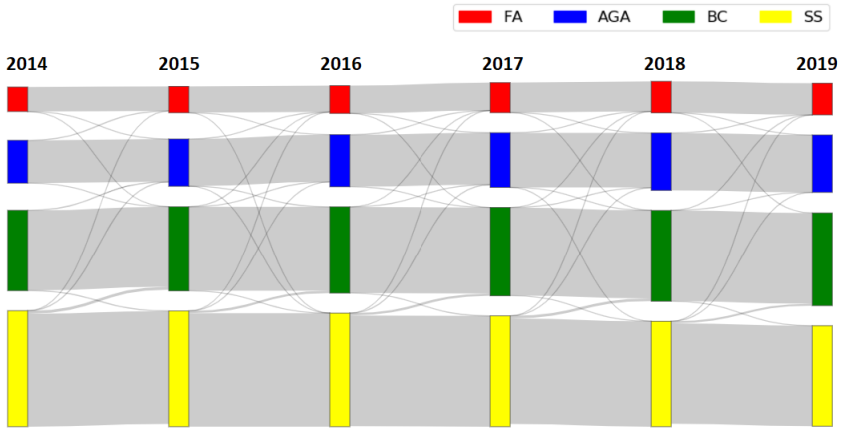


Figure 3.5: Temporal illustration of intra and inter class quality changes over the years 2014-2019.

3.3.5 Cyclic switch of qualities

Let us assume an article P is assigned to the following quality classes at different timestamps (in an increasing timeline)

$$\textbf{class1} \Rightarrow class2 \Rightarrow class3 \Rightarrow \textbf{class1} \tag{3.2}$$

and, recursively,

$$\textbf{class1} \Rightarrow (classX)^{+} \Rightarrow \textbf{class1} \tag{3.3}$$

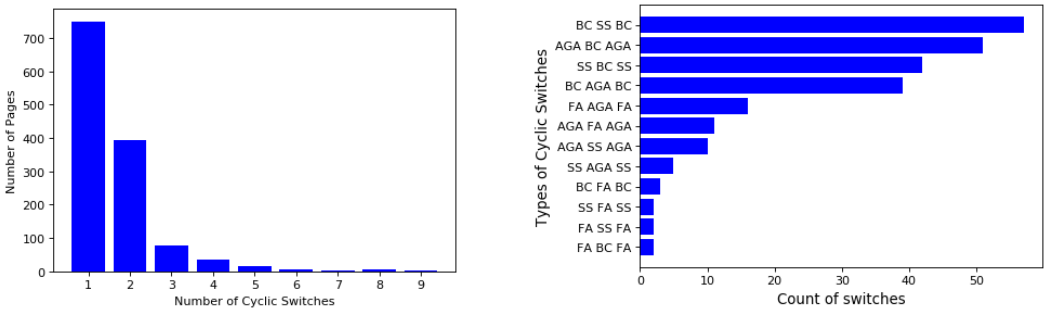
According to equation 3.2, the *class1* can be any among the four types of quality classes and *class2*, *class3* can include any permutation of remaining three classes (P_2^3). We impose the constraint that the intermediate consecutive class levels can not belong to the same class, for example, *class2* and *class3* cannot be same. We denote this type of temporal pattern of quality assignment as *cyclic switches* and in equation 3.2, the length of the cyclic switch is 4. Such cyclic switches may at times be a possible outcome of editorial conflicts as opposed to organic quality shifts.

Quality switch war: We find 1286 articles that exhibit one or more cyclic switches. A majority of the articles have switches of length three (1258 articles in all)[21]. The histogram in Figure 3.6a presents the distribution of articles that undergo different number of cyclic switches. As expected the distribution decays very fast. The average turnaround time for cyclic switches in 1286 articles (e.g., length three switches which constitutes the majority) is observed to be 920.89 days. At the same time, we observe 180 pages out of the 1286 articles contain cycle switches in which the minimum duration of turnaround time is less than 15 days. We therefore investigate the distribution of these *very rapid* cyclic switches in Figure 3.6b. In addition, we also saw that wikibots[22] are responsible for cyclic changes in 0.05% cases only compared to the humans. These results together point to the fact that an article experiencing multiple cyclic switches often with short turnaround times in its entire lifespan is an extremely non-trivial behaviour. Likewise edit war this behaviour can be attributed to *quality switch war* that arises possibly due to the continuous

[21] From the unclubbed to the clubbed version, $\sim 90\%$ of the switches are retained.

[22] https://en.wikipedia.org/wiki/Wikipedia:Bots

conflicts among the editors regarding their perception about the quality of that article.



(a) Distribution of articles undergoing different number of cyclic switches.

(b) Distribution of length three cyclic switches with turnaround time less than 15 days.

Figure 3.6: Different distributions of cyclic switches.

Prevalent switches: Further analysis, tabulated in Table 3.6 shows that the count of cyclic switches decreases exponentially with the increase in the length of the switch. We also find that the following cyclic switches <**BC**, **SS**, **BC**>, <**SS**, **BC**, **SS**>, <**BC**, **AGA**, **BC**>, <**AGA**, **BC**, **AGA**> are the top four (in that order) length three switches in terms of occurrence across different articles. This indicates that typically low quality articles are vulnerable to more frequent switches.

Long and short cyclic switches: Based on a closer inspection into the talk pages of some of the articles, we find that cyclic quality switches can be of different lengths. For example, we observe instances of cyclic switches [FA→ FA] of varying lengths, i.e., three and eight respectively within a single article (see Figure 3.7).

Length	3	4	5	6	7	8	9
Count	1993	70	32	4	6	1	1

Table 3.6: Count of cyclic switches of varying length of switches.

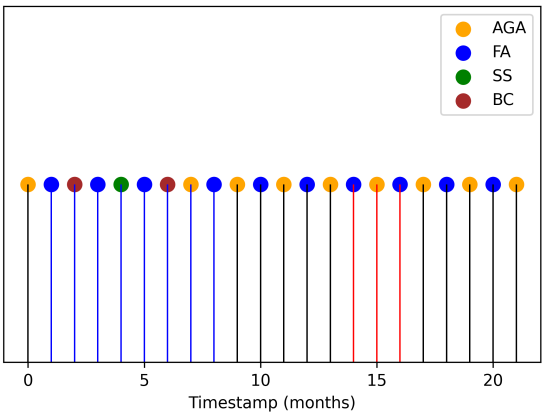


Figure 3.7: Qualities assigned to an article at different time points. The region plotted in blue lines indicate a cyclic switch [FA → FA] with larger number of state changes to get back to the initial state compared to the region indicated by the red lines. The black lines indicate the assigned quality at a particular time but do not correspond to a switching behaviour.

3.4 Detection of Quality Change Points

The extensive analysis in the previous sections show that quality assessments of Wikipedia pages experience various hindrances over their lifespan. These range from frequent and unpredictable switches of quality to absolute stagnancy. Therefore, an early alert system that can provide timely signals to the editors suggesting the requirement of quality assessment of an article is very much needed. In this section we attempt to bridge this gap by building an automated unsupervised feature-based approach to discern the change points of article quality as early as possible which will be the first step to developing such an alert system.

3.4.1 The quality indicators

We carry a detailed analysis of the typical factors through which article quality evolves over time and based on these we attempt to formulate very intuitive categorical features

that could be responsible for such changes. The features that we selected can be grouped into three classes.

- Contribution based features (editors' participation attributes) [*aka* G_c]
- Activity based features (edit pattern attributes) [*aka* G_a]
- Content based features (article's attributes) [*aka* G_p]

For extracting *contribution based features* (G_c), we parse every revision of the articles (an xml file containing all the revisions of an article from the date of the creation) in our dataset and collect the editors' usernames editing the revisions. In case of unregistered editors, the contributor's username is mentioned as anonymous IP address. Similarly, we collect the editors' information from the revision of talk pages of articles individually. The count of the editors, such as distinct registered (unregistered) editors editing the articles and talk pages are used as the features. For extracting *activity based features* (G_a), the number of revisions an article and its talk page passed through (calculated per month and week basis), mean and variance of time difference between two consecutive revisions of both the article page and talk page at the granularity of months are calculated. Similar to the features under (G_c), we parse the revisions (i.e., xml files) of article pages and talk pages to collect the activity related features. In case of *content based features* (G_p), we extract the content of the latest revision of the articles in every month and compute features of this category. We use the mediawiki[23] parser to parse the wiki text and computed various content based features. In the following we present an elaborate description of all these features.

Contribution based features (G_c)

The features we include in this category describe the involvement of the editors over the revisions of an article. The choice of these features is motivated by the fact that the

[23]https://pypi.org/project/mwparserfromhell/

contribution of editors should have some impact on the quality change of an article [103]. Our hypothesis is that the temporal changes in the number of editors contributing to an article could bear early signs of quality shift.

- *The number of distinct registered editors editing the talk pages* (**F1**) : Talk pages act like a discussion forum among the editors regarding the content, organization of the specific article and the participation of editors in discussions typically get enhanced close to points of quality change.
- *The number of newly added registered editors editing talk pages* (**F2**) : Similar to the previous feature, we consider the count of the new editors commenting in the talk pages as one of the contribution based features.
- *The number of distinct unregistered editors editing the talk pages* (**F3**) : Once again, our hypothesis is that a large number of unregistered editors (anonymous IPs) participating in the talk page discussions of an article at a point should be indicative of a quality change in the near future.
- *The number of distinct registered editors editing the article main page* (**F4**) : The increase/decrease in the number of editors, editing the main content should impact quality change.
- *The number of newly added registered editors editing the main page* (**F5**) : We use the count of new editors editing an article (suggestive of enhanced attention because of the ongoing review process or a sudden increase in popularity of the respective article) as a quality change indicator.
- *The number of unregistered editors editing the article main page* (**F6**) : A sudden change in the number of unregistered editors (anonymous IPs), editing the main content should be indicative of quality change.

We do not consider the editors' experience or their level of engagement in our contribution based features to reduce the noise in representing the pages individually.

Activity based features (G_a)

Earlier research [197] shows that there exists a strong correlation between article quality and the editing activity; an abrupt spike in activity, or very less activity could be potential indicators of quality change. We include various edit activity based features in our model.

- *Mean time elapsed (at the granularity of months) between two consecutive revisions of the article main pages.* (**F7**)
- *Variance of time elapsed (at the granularity of months) between two consecutive revisions of the article main pages.* (**F8**)
- *Mean time elapsed (at the granularity of months) between two consecutive revisions of the article talk pages.* (**F9**)
- *Variance of time elapsed (at the granularity of months) between two consecutive revisions of the article talk pages.* (**F10**)
- *Number of revisions of the talk pages (at the granularity of months).* (**F11**)
- *Number of revisions of the talk page (at the granularity of weeks).* (**F12**)
- *Number of revisions of the main page (at the granularity of months).* (**F13**)
- *Number of revisions of the main page (at the granularity of weeks).* (**F14**)

Content based features (G_p)

Previous works show that article content plays an important role in assessing quality and hence we consider the article text, i.e., content of the main namespace as possible indicators of quality change. In particular, we extract the following features for every page as reported in [175, 207].

- *Article length in bytes.* (**F15**)

- *Number of references.* (**F16**)
- *Number of categories mentioned in the text.* (**F17**)
- *Number of links to other articles.* (**F18**)
- *Number of citation templates.* (**F19**)
- *Number of non-citation templates.* (**F20**)
- *Number of images/article length.* (**F21**)
- *If infobox template exists.* (**F22**)
- *Number of level 2 section headings.* (**F23**)
- *Number of level 3+ section headings.* (**F24**)
- *Information noise score.* (**F25**) (adapted from Zhu and Gauch [215])
- *Readability scores.* (**F26 : F34**) : 9 types of readability scores.

Readability measures how interpretable an article is to the readers and computes a score based on the use of language in the article. This is used as an important measure to reflect the encyclopedic standard of representing its content. We include 9 readability scores in our features as discussed in [175].

Finally, every revision of an article is represented as a vector of all the above 34 features for the detection of quality change points.

Correlation among the features

To examine the relationship among the different features (G_c , G_a, G_p), we compute the Pearson's correlation coefficients between all possible feature combinations. Unlike the standard form of feature vectors used in prediction models, every feature in our work is represented as the time series of distinct feature values for the time span (i.e., 222

months) we consider. This is true for every individual article. Formally, let us assume, a feature F_i of an article A_i is represented as time-series of m timestamps, i.e., spanning over m months. Also, let us denote the i^{th} timestamp as T_i for the feature F_i. We have 14872 articles (i.e., say n) in our dataset that undergo quality changes at least once in the lifetime. We average the feature F_i of n articles at timestamp T_i. We follow the same method for all the timestamps (which is 222 in number) and thus obtain 222 values of F_i. The above mentioned procedure is followed for all the features. This results in a matrix of dimension of 34 * 222, in which the number of features and timestamps is 34 and 222 respectively. Figure 3.8 shows the correlation coefficient matrix for the 34 features.

We observe that the features belonging to same category (G_c or G_a or G_p) are highly correlated but are less correlated with the features of other categories. However almost all the coefficients are in the range 0.2 to 0.9 except for the pairs which belong to the same category of features and are semantically close to each other. For example, the features - number of editors editing the article pages (i.e., F4) and number of new editors editing the article pages (i.e., F5) under *contribution based features* (G_c) or the different readability scores except the feature F26 of *content based features* (G_p) are closely related and hence are showing highly positive correlation in the heatmap. Also, we find that *content based features* (G_p) except readability scores are less correlated with each other, especially the feature whether infobox template exists (i.e., F22) is almost zero correlated within the group.

Analysis of temporal changes in the features

We present a number of quality change patterns and their statistics so far. As a next step we want to observe the temporal pattern of changes of the features at the point of the quality change. In particular we are interested in cases that undergo both *promotion and demotion in quality*, which result in a selection of 54 pages that undergo a change from the FA quality (the highest quality) to some lower quality (AGA, BC, SS) at least once in their life-time. The time window is set as the 24 consecutive timestamps (i.e., months) including the change point. The articles that earn the FA quality, undergo several rounds

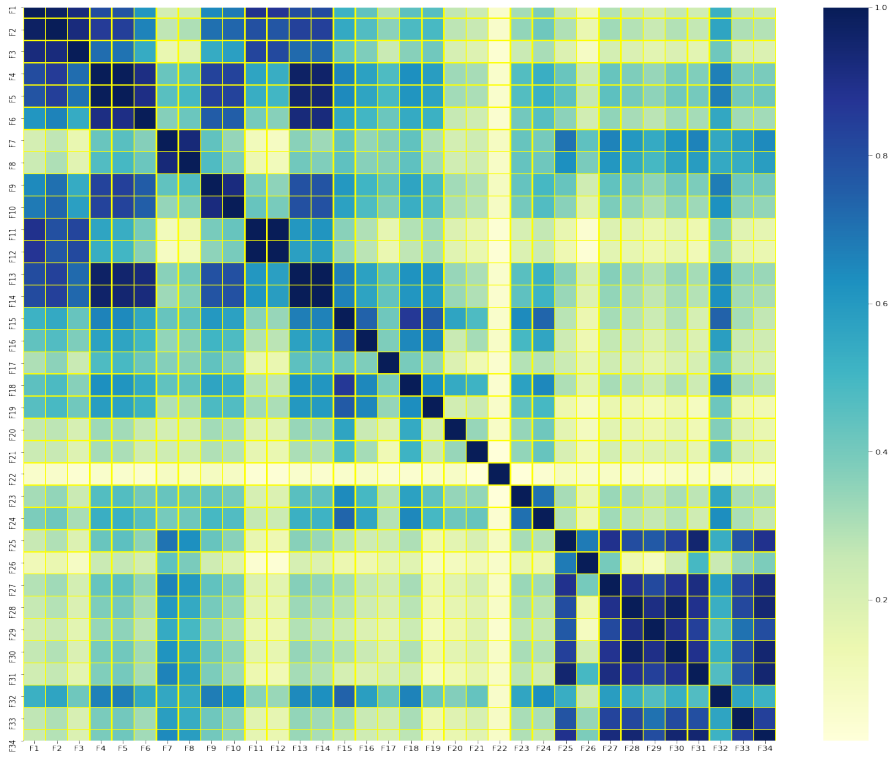


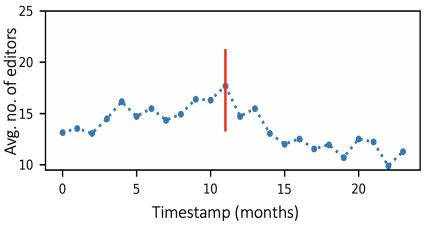
Figure 3.8: Heatmap showing the correlations between the various categories of features. Blue → Sky → White indicates 1.0 → 0.5 → 0.0 correlation values. The heatmap follows the feature order (top to bottom) in which the first 6 features denote the *contribution based features* followed by 8 features of the *activity based features*. The last 20 features in the heatmap define the *content based features*.

of review and hence abrupt changes in the feature space are expected before and after the quality change (e.g., change from FA to BC). The red line indicates the change point in each of the plots in Figure 3.9. In every case, the data points represent the mean values of the features of 54 pages.

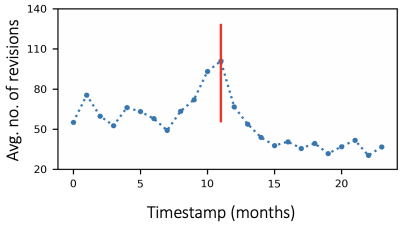
- As shown in Figure 3.9(a), the number of distinct registered editors editing the article main page declines suddenly after the quality change point. If we consider 4 consecutive timestamps on either side of the change point (red line), the change is sharp on both sides. This indicates that a large number of editors are active when the pages reach the FA status but many of them stop editing (the main page) after the demotion in quality.

- Similarly in Figure 3.9(b), we plot the number of revisions of the main page and observed very similar trends as with the number of editors. There is a huge difference in the number of revisions on both sides of the red line, indicating a sharp change in the activity of the community.
- In Figure 3.9(c), the length of the article increases sharply at the point of quality change. The increase in bytes can signal that the editors add more content to the articles after the demotion compared to the earlier phase.

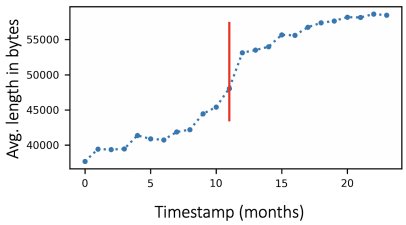
Overall, the temporal patterns of the three features show shift in their mean values at the point of quality changes. The analysis reveal that such patterns could be particularly helpful for the CPD methods to detect the change points.



(a) Number of distinct registered editors editing the article main page



(b) Number of revisions of the main page



(c) Article length in bytes

Figure 3.9: The temporal pattern of three features from the three feature categories. The features are computed for 54 pages that have changed from the FA quality to some lower quality and the red line indicates the change point. The data points represent the mean values of the 54 pages in each case.

3.4.2 Change point detection

The time-series settings

We now briefly describe the problem of *change point detection* (*CPD*) and its resemblance to the quality change detection in our setting. In a CPD problem, given a set of observations $\mathcal{Y}$ for time steps $t = 1, 2, \ldots$, we need to identify the time indexes where there is an abrupt change in the behavior of the time series, specifying a probable signal to the alteration of the data generation process. These indexes are denoted as *change points*, and a number of CPD algorithms are used to detect the change points. The algorithms are about to estimate the unknown instances where the characteristics of the series change abruptly. Depending on the context, the change points are used in evaluating the correctness of the algorithm. In our settings of quality change point detection, we postulate the analogy as follows.

- Individual article is assumed as a separate time series, and the timeline is expressed in months. In other words, we consider the revisions on a monthly basis.
- The sample data points are the latest revision in each month given a typical Wikipedia article. Since quality assessments are unlikely to change with every revision or timestamp, assigning quality ratings to each timestamp would render the time series, i.e., the feature set, redundant. In addition, quality switches within a short period, often caused by "quality switch wars," can introduce substantial noise into the dataset. To address this, we use the quality ratings from the last revision of each month, which allows us to retain 95% of the ground truth quality change points.
- The change points are the time instants, specifically the months where the quality of the article was changed. We shall be using the terms months and timestamps interchangeably.

Mathematically, let $\mathbf{y}_t \in \mathcal{Y}$ denote the representation of the article at the timestamps

T_t $\forall t = 1, 2, \ldots, N$. Here, N is the number of months in our timeline. Let us denote the size of the domain of every $\mathbf{y}_t$ by *d*, which is the total number of features described earlier, and therefore $\mathcal{Y} \subset \mathrm{R}^d$. We denote $\mathbf{y}_{i:j}$, $i < j$ as the segment of timestamps $T_i, T_{i+1}, \ldots, T_j$. The timestamp ordered set of quality change points is denoted by $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$, where n is the number of quality change points for the given article and used as the ground truth. We use $q_0 = T_1$, the timestamp at which the page is created, and $q_{n+1} = T_N + 1$ as the endpoints of the quality change point set. It is required to note that we add these points only for convenience of evaluation metrics which are defined later in this section, and we do not consider these points as the actual quality change points.

The CPD algorithms

We experiment with a few CPD algorithms ranging from one of the earliest proposed models to the more recently proposed ones. Since our feature space is multi-dimensional, we experiment with the multivariate setting only. Furthermore, since our objective is to analyze and infer observations from the features that are used in detecting the quality change points, we experiment mainly with the offline CPD algorithms for retrospective detection and analysis of the change points. In the offline multivariate change point detection algorithms, the general objective is to optimize the following cost function

$$\min_{\mathcal{Q}} \sum_{i=1}^{n+1} \mathcal{L}(\mathbf{y}_{\mathcal{Q}_{i-1}:\mathcal{Q}_i-1}) + \lambda P(n) \tag{3.4}$$

where $\mathcal{L}(.)$ is the loss associated for the segment $\mathbf{y}_{\mathcal{Q}_{i-1}:\mathcal{Q}_i-1}$, λ is a hyperparameter, and $P(n)$ is a penalty on the number of change points. The intuition for this function is to minimize the loss in grouping the timestamps with the same quality of the article in a single contiguous segment. Also note that the point at which the quality changes ($\mathcal{Q}_i$) is considered in the new segment. For the close compatibility with change point detection and localization in multivariate data with multiple change points, we apply the following change point detection algorithms in our work.

- Binary Segmentation [*aka* **BINSEG**] - It is one of the earliest methods for detecting

the change points that greedily splits the timestamp series into disjoint segments based on optimising a predefined cost function. For the quality change detection, we use the cost function defined in Equation 3.4. The time complexity associated with this method is $O(NlogN)$.

- Pruned Exact Linear Time [*aka* **PELT**] - This is an offline method that works through minimising the aforementioned cost function over possible numbers and locations of change points. Through minimization, the approach achieves the optimal number and location of change points that has a computational cost which under mild conditions, is linear time in the number of observations.

- Non-parametric Change Point Detection [*aka* **ECP**] - It is a nonparametric approach for detecting the change points. For a set of multivariate observations of arbitrary dimensions, the model performs a nonparametric estimation of both the number of change points and the locations at which they occur. The estimation of the change points is based on hierarchical clustering of the timestamps.

3.4.3 Evaluation

For evaluating the performance of the methods, we use the metrics described in [184]. These metrics are compatible with multiple change point (ground truth) setting and also quantify the consistency of the annotations of the timestamps. These metrics can be roughly categorised into clustering and classification metrics. The locations of the ground truth change points are denoted by the ordered set $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}\ \forall g_i \in \{T_1, T_2, \ldots, T_N\}$ and $g_i < g_j$ for $i < j$. The set $\mathcal{G}$ partitions the timestamps into disjoint sets $s_j \in \mathcal{S}$, where s_j is the segment from T_{j-1} to $T_j - 1\ \forall j \in \{1, 2, \ldots, k+1\}$. The clustering based metrics evaluates the CPD algorithms based on the view that change point detection inherently aims to divide the timestamps into distinct regions with a constant quality of the article.

Change point evaluation as clustering: Among the different clustering metrics proposed by several algorithms [16, 59, 77, 88], we use the *covering metric* because of its

ability to show the true performance of methods that report many false positives. For any two sets s, s', the Jaccard index is computed using the following expression

$$J(s, s') = \frac{|s \cap s'|}{|s \cup s'|} \tag{3.5}$$

The authors in [16] define the covering metric of a partition $\mathcal{S}$ by partition $\mathcal{S}'$ as

$$\mathcal{C}(\mathcal{S}, \mathcal{S}') = \frac{1}{N} \sum_{s \in \mathcal{S}} |s|. \min_{s' \in \mathcal{S}'} J(s, s') \tag{3.6}$$

where partition $\mathcal{S}$ is the partition induced by the ground truth set $\mathcal{G}$, and $\mathcal{S}'$ is the partition induced by the set $\mathcal{Q}$ predicted by the model.

Change point evaluation as classification: A different set of evaluation metrics for CPD algorithms considers the change point detection as a classification problem between the "change point" and "non-change point" classes [8, 101]. The simple metrics such as accuracy will be highly skewed because the number of quality change points of a page will be very small compared to the total number of revision timestamps of the article. Therefore, we look at the effectiveness of the algorithms in terms of *precision* and *recall*. The set of true positives in the predicted set $\mathcal{Q}$, denoted by $\mathrm{TP}(\mathcal{G}, \mathcal{Q})$, consists of all the timestamps $g \in \mathcal{G}$ for which $\exists q \in \mathcal{Q}$ such that $|g - q| \leq M$, while ensuring that only one $q \in \mathcal{Q}$ is used for one $g \in \mathcal{G}$. The value M is a commonly defined margin of error around the true change point location to allow for minor discrepancies, which is an usual practice in evaluating the change point detection algorithms [101, 183, 184]. However, the additional condition imposed avoids double counting, so that among the multiple detection within the margin around a true change point only one is recorded as a true positive [101]. The precision and recall are then defined as follows.

$$P = \frac{|\mathrm{TP}(\mathcal{G}, \mathcal{Q})|}{|\mathcal{Q}|} \tag{3.7}$$

$$R = \frac{|\mathrm{TP}(\mathcal{G}, \mathcal{Q})|}{|\mathcal{G}|} \tag{3.8}$$

According to this definition the false positives are the ones that do not have any corresponding ground truth points.

3.5 Experiment and Results

We run the three algorithms – **BINSEG**, **PELT** and **ECP** on a sample of our dataset (discussed in the next section) for all possible combinations of the mentioned features. In addition, we also present an ensemble of the three algorithms, which we term as the **HYBRID** algorithm. This simply takes the best of the three methods and outputs the same. Further, inspired by the feature correlation analysis in the previous section, we present a number of ablation studies to understand which features are the strongest determinants of quality change points.

3.5.1 Experimental setup

For preparing the sample set to run the algorithms, we consider only those articles that have more than one ground truth quality change points. This leaves us back 14872 out of 30826 articles. This is since it is essential for the CPD methods to have at least one ground truth point to distinguish the change in data generation process of the time series. Every article was represented as a distinct time series of revisions at different timestamps, and each revision as a d dimensional feature vector. We consider only the latest revision in a month to filter out the noise in the dataset that occurs due to the quality switch war, which has been explained before (cf section 3.3.5). Moreover, we experiment with various time intervals ranging from ten days to six months and observe that by using one month, we retain 95% of the quality change points, which also produces the best results. Further, we consider that revisions in the main namespace (namespace 0) should always occur before those in the talk namespace (namespace 1). This ensures that the quality assessment timestamp is always later than the revisions containing modifications made to the article content. By following this approach, we guarantee that editors take into account all necessary changes in the article before conducting their assessments; we proceed with these revisions only. Since all features we use are cumulative in nature, this choice does not affect the feature computation process. To investigate the performance of CPD algorithms, we divide the dataset comprising of 14872 articles into two sample sets – train and

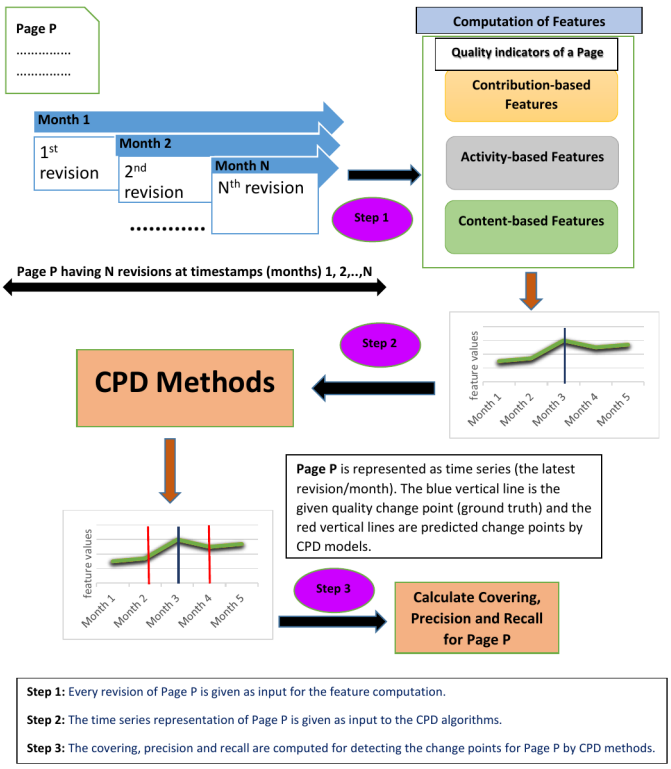


Figure 3.10: The pipeline for unsupervised change point detection for an individual article.

test sets. We divide the dataset comprising 14872 articles into two sample sets - training and test sets. The train-test split for our experiment is set to 80:20. We try to maintain a uniform distribution of articles of different quality classes across the two sets so that the results do not get dominated by the performance of a single class containing a specific set of articles. To achieve the best results, we tune the hyperparameters of the CPD algorithms on the training set and later tested them on the test set. A schematic diagram of our experimental setup for the change point detection is illustrated in Figure 3.10.

3.5.2 Hyperparameter settings

The penalty value, i.e, *pen_val* of the PELT and *n_bkps* of the BINSEG algorithm, which determines the significance of the change points identified, is set to 1, which perform best in the range of values from 1 to 8. The mean number of change points (i.e., ground truth) observed for the articles in the training set is 2.63. For both PELT and BINSEG methods, we use rbf to model the cost function. For the ECP algorithm, we vary the *min_size* parameter, which defines the minimum gap between two successive change points from the set of values 2, 5, 10, 15, 20. The best results are observed at 5. The optimal value of the hyperparameters are observed for the articles in the training set. For evaluation in terms of precision and recall, the margin of error M is set to 5 ($\pm$ 5 on either side of the predicted point). We observe the mean value of the interval between two consecutive change points (i.e., ground truth) to be 30.35 months, which is considerably higher than the error margin we use in our experiments. We ran the three CPD algorithms with their optimal hyperparameters on every test page individually and finally aggregated metric values over all the test samples. For the HYBRID method, we take the maximum one reported by the three algorithms in their best hyperparameter settings. The mean values for each of the metrics, i.e., covering, precision, recall for all the combinations of the features are reported in the Table 3.7.

3.5.3 Key results

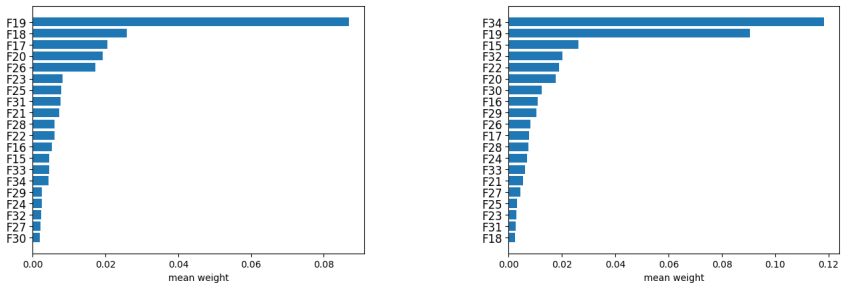
We achieve the best covering of **0.76** from BINSEG. The best result for ECP and PELT are 0.6. As expected, the content based features alone are sufficient to produce the best results (see [131] for similar observations). The highest achievable precision and recall are **0.6** (BINSEG) and **0.61** (PELT) respectively. Once again these best results are achieved using just the content based features. Overall among the three algorithms, PELT achieves the best compromise considering all the three results (covering $= 0.6$, precision $= 0.37$ and recall $= 0.61$). The HYBRID method, as it should be, achieves the best covering (0.79), precision (0.68) and recall (0.69) for the content-based features.

To gain deeper insights into the individual functioning of CPD algorithms, we conduct the following experiment. We select articles where the coverage determined by BINSEG exceeds that of ECP and PELT for the content-based feature combination (G_p). Our goal is to examine the specific cases where BINSEG outperforms the other algorithms. Since time series analysis evaluates each time point independently, we restrict the experiment to the first ground truth change point in the selected articles. We further refine our selection by filtering out articles where only BINSEG correctly detects the ground truth change point while the other two algorithms fail. This allows us to formulate a classification problem: class 1 represents successful change point detection by BINSEG, whereas class 0 includes instances where both ECP and PELT fail. We then utilize content-based features from the articles in both classes for the specific change point and classify them using a Random Forest classifier. Next, we try to explain the predictions by applying LIME analysis to the test set instances to identify the most influential features in the classification process, as illustrated in Figure 3.11a. The features F19, F18, F17, F20, F26 under content-based features play a crucial role in detecting change points by BINSEG. A similar approach is applied to PELT, for instances where it detects change points more accurately than BINSEG and ECP. Among the content-based features, F34 plays a significant role in PELT's performance, as shown in Figure 3.11b. Like BINSEG, Feature F19 (i.e., the number of citation templates) is also important, though it has lesser significance as per the LIME weights. This weight analysis process can be extended to other feature combinations to further investigate their impact on the performance of CPD algorithms—BINSEG, ECP, and PELT. Thus, different feature combinations enhance the effectiveness of the individual algorithms demonstrating the usefulness of ensembling.

3.5.4 Additional experiments

Further, we analyze the obtained results based on two special criteria.

- **Criteria 1:** We consider only those articles that have at least three change points (i.e., assuming that larger number of change points on the time series should help



(a) feature importance: BINSEG algorithm

(b) feature importance: PELT algorithm

Figure 3.11: Feature importance as explained by LIME for the change point instances where BINSEG and PELT outperform other CPD algorithms.

Features	BINSEG [n_bkps = 1]			ECP [min_size = 5]			PELT [pen_val = 1]			HYBRID		
	Covering	Precision	Recall	Covering	Precision	Recall	Covering	Precision	Recall	Covering	Precision	Recall
G_c	0.68	0.36	0.29	0.39	0.23	0.27	0.52	0.28	0.47	0.73	0.51	0.60
G_a	0.68	0.37	0.30	0.41	0.27	0.29	0.52	0.29	0.47	0.74	0.53	0.59
G_p	0.76	0.60	0.46	0.60	0.45	0.45	0.60	0.37	0.61	0.79	0.68	0.69
$G_c \oplus G_a$	0.68	0.38	0.30	0.42	0.27	0.31	0.53	0.29	0.45	0.74	0.53	0.59
$G_a \oplus G_p$	0.75	0.56	0.43	0.59	0.42	0.44	0.60	0.37	0.58	0.78	0.66	0.68
$G_p \oplus G_c$	0.74	0.53	0.41	0.56	0.39	0.43	0.60	0.37	0.57	0.78	0.64	0.60
$G_c \oplus G_a \oplus G_p$	0.74	0.53	0.41	0.56	0.39	0.43	0.60	0.37	0.57	0.78	0.64	0.67

Table 3.7: CPD outcome: A comparison of the BINSEG, ECP and PELT algorithms on test set. Best results are highlighted in green. Results highlighted in blue are the best among those achieved by the **HYBRID** method.

in having a better inference). This makes 1982 articles in this category.

- **Criteria 2:** Here we consider only those articles whose latest class is FA (3190 articles). This is precisely to understand how well the model performs for the highest quality and possibly the most important class.

We run the ECP algorithm on the test samples belonging to the mentioned criteria 1 and criteria 2 (other algorithms produce similar results and hence, not shown) and obtain the values of the three evaluation metrics. The train-test split (80:20) and hyperparameter values remain same as described earlier. The results are noted in the Table 3.8. As is expected, the results are considerably better than the entire dataset for all the feature combinations.

features	Criteria 1			Criteria 2		
	Covering	Precision	Recall	Covering	Precision	Recall
G_c	0.53	0.39	0.33	0.52	0.37	0.37
G_a	0.58	0.48	0.40	0.56	0.44	0.44
G_p	0.72	0.67	0.47	0.71	0.65	0.56
$G_c \oplus G_a$	0.59	0.48	0.41	0.57	0.43	0.44
$G_a \oplus G_p$	0.71	0.60	0.50	0.69	0.58	0.55
$G_p \oplus G_c$	0.69	0.58	0.46	0.67	0.55	0.53
$G_c \oplus G_a \oplus G_p$	0.67	0.56	0.48	0.65	0.52	0.53

Table 3.8: CPD outcome of ECP algorithm (other algorithms show similar trends and hence not shown) for the two different special criteria. Best results for each criteria are highlighted in green.

3.5.5 Ablation study

To understand the contribution of specific features to the outcome of the CPD algorithms, we perform several ablation studies of feature combinations of three specified categories - G_c, G_a and G_p. The results show that content based features G_p achieves the best performance. We perform an ablation study in which one feature is masked and the CPD algorithm is run for the remaining features. In every step the drop in the values of three metrics are noted. The drop is not significant for masking any single feature. We report the top 3 changes in values for masking three features in the Table 3.9. While we show the results for PELT (with *pen_val* set to 1), the other algorithms show similar trends.

Feature	PELT [pen_val = 1]		
	Covering	Precision	Recall
$G_c - F1$	0.51	0.27	0.45
$G_a - F14$	0.53	0.28	0.45
$G_p - F32$	0.58	0.34	0.55

Table 3.9: CPD (PELT) Outcome: Results for masking three features individually. Three features that are selected from each category of G_c, G_a, G_p are (i) Number of registered editors editing talk pages, (ii) Number of revisions of article page per week and (iii) difficult words (readability score) respectively.

Next, we try combination of different features irrespective of the feature category

and run the CPD algorithm. The choice of the features are motivated by the correlation coefficients as shown in the heatmap (see Figure 3.8). We use the abbreviations for the combination of features as mentioned below. The notations of the features are same as in the heatmap (see Figure 3.8).

- **G1:** All the readability features, i.e., F26 : F34
- **G2:** All the content based features, i.e., F15 to F25 except the readability features.
- **G3:** All the features of G2 and the readability feature F32.
- **G4:** A subset of content based features, F15 to F21 and the features, F23 and F24.
- **G5:** All the features of G4 and the readability feature F32.
- **G6:** A subset of activity features, F9 to F14.
- **G7:** The features mentioned in G6 and the readability feature F32.
- **G8:** The features of G7 and all the contribution based features.

We follow the similar train-test split in tuning the hyperparameters and the result is tabulated in the Table 3.10. Once again we use PELT with the hyperparameter *pen_val* set to 1. While in Table 3.7 we observe that the best performance comes from the content based features when we dive deeper we observe that a selected combination G7 of activity and content based features chosen as per the heatmap in Table 3.9 gives superior performance. In specific the features that play instrumental role are

- the mean and the variance of time elapsed between two consecutive revisions of the article talk pages (at the granularity of months),
- the number of revisions of the article talk pages (at the granularity of both months and weeks),
- the number of revisions of article main pages (both at the granularity of months and weeks), and.

- presence of "difficult words" – a word is considered "difficult" if it does not appear in a list of 3,000 common English words that a fourth-grade American student can typically understand. In our dataset, we identified several examples of highly frequent *difficult* words, including "xenon," "pipeline," "anole," "touchdown," "epilepsy," "carfilzomib," etc.

Further we see that G3 and G5 feature groups (i.e., the content features article length (in bytes), number of references, number of categories mentioned in the text, number of links to other articles, number of citation templates, number of images/article length, whether infobox template exists, number of level 2 section headings, number of level 3+ section headings and the anti-correlated readability feature "difficult words" (among all the readability features) together brings the highest gain in covering. To summarize a set of judiciously selected feature combinations (based on the feature correlations), e.g., mean and variance of time elapsed between two consecutive revisions and the number of revisions of the article talk pages which typically correspond to *the organisational attributes* of the peer-production system and certain *readability attributes* like the presence of "difficult words" become crucial when we perform an in-depth analysis which do not manifest in the aggregate level results (i.e., in Table 3.7).

	PELT [pen_val = 1]		
Features	Covering	Precision	Recall
G1	0.53	0.32	0.74
G2	0.52	0.28	0.37
G3	0.61	0.39	0.57
G4	0.52	0.27	0.37
G5	0.62	0.39	0.57
G6	0.53	0.30	0.51
G7	0.62	0.41	0.69
G8	0.60	0.39	0.63

Table 3.10: CPD (PELT) outcome: Results for different combination of features on test data. Best results are highlighted in green.

In order to further understand the classwise importance of features, we report the performance metrics for each class in Table 3.11. This based on the feature groups

introduced earlier (i.e., G1–G7) and the PELT model. An universal trend is that the best covering is obtained for the feature group G5. Overall, for the majority of the performance metrics, G7 is the winner thus pointing to the universality and robustness of our results.

	FA class			AGA class			BC and SS class		
Features	Covering	Precision	Recall	Covering	Precision	Recall	Covering	Precision	Recall
G1	0.56	0.38	0.77	0.56	0.37	0.80	0.49	0.25	0.70
G2	0.61	0.39	0.48	0.54	0.32	0.41	0.47	0.17	0.28
G3	0.61	0.39	0.48	0.54	0.32	0.41	0.55	0.25	0.44
G4	0.61	0.39	0.48	0.54	0.32	0.41	0.47	0.17	0.28
G5	0.70	0.51	0.67	0.65	0.47	0.65	0.55	0.25	0.44
G6	0.59	0.39	0.60	0.55	0.34	0.54	0.46	0.18	0.41
G7	0.69	0.52	0.76	0.65	0.48	0.75	0.55	0.27	0.56
G8	0.66	0.49	0.73	0.63	0.45	0.69	0.53	0.25	0.51

Table 3.11: CPD outcome: A comparison of different combination of features on different quality articles. Best results are highlighted in green.

3.5.6 Baseline: ORES

Likewise the popular social media platforms, Facebook, YouTube, Twitter and other corporate and government organizations, Wikipedia has a significant number of AI-based resources that help the community to take decisions at large scale. Among the pool of bots, human-in-the-loop assisted tools, expert systems, ORES[24] is a web-service or API, designed for Wikimedia projects to provide an automated solution to critical wiki-works, for example, predicting edit quality, article quality etc. ORES is trained with a large number of machine learning classifiers, operating in real time on the Wikimedia foundation's backend servers and can output a quality score for the given edit or page as a query to the service. In this section we compare the performance of ORES with our CPD algorithms.

The experiment: For a typical article, we provide every revision to ORES as query and ORES predict one of the six quality classes. If the predicted class is different from the previous revision, we mark the revision as a predicted quality change point. In such a way, we get the predicted outputs for every revision of a page which is analogous to the

[24]https://www.mediawiki.org/wiki/ORES

outcomes of CPD algorithms. Now, to compute the precision, recall and covering, we assume the margin of error M same as that of the CPD predictions. Similar to our method, if the predicted change point from ORES is within the margin of error of the ground truth (the change points marked by the editors in the talk pages), the point is marked as a true positive. This allow us to compute all the metrics – covering, precision and recall in this setting. As we shall see, the quality indicators (features) and the classification strategy (CPD algorithms vs machine learning classifiers) used in the two settings produced the difference of correctness in prediction task.

Result: We perform the comparison experiments for the two subsets of pages and the results are tabulated in the Table 3.12. In both cases, we compare the HYBRID method of CPD with ORES. The optimal hyperparameter values remain same as decided earlier.

- Set 1: Set of pages that have at least one change point in the ground truth reality.
- Set 2: The set articles whose latest class is assigned to FA quality.

We observe that our HYBRID CPD algorithm by far outperforms the results obtained from ORES that works on an ensemble of machine learning classifiers.

	CPD HYBRID method			ORES		
#Articles	Covering	Precision	Recall	Covering	Precision	Recall
Set 1	0.79	0.68	0.69	0.56	0.31	0.60
Set 2	0.82	0.80	0.73	0.63	0.40	0.71

Table 3.12: Performance comparison of HYBRID method with ORES on the test split.

3.6 Summary

In this chapter, we inspect the life cycle of article quality over a massive dataset of $30k$ articles and notice varying dynamical patterns of quality moderation. One of the most

important findings of our work is that while more than 50% articles do not experience any quality change over their entire lifespan, there are some articles that undergo quality shifts multiple times and that too within very short spans of time. We also observe possibilities of quality switch wars apparent from the rapid cyclic switches in various Wikipedia articles. As a second objective, we model the detection of quality changes as a change point detection problem. We leverage the diverse collaborative features of Wikipedia and feed them to standard change point detection algorithms. In general, it is true that content-based features are one of the most predictive. However, as we deep dive into the feature space, we observe that certain combinations of content and activity features show competitive (and, in fact, better) performance. In particular, the presence of "difficult words" (a readability feature) in the article seems to be a strong indicator of quality change. Our findings align with those of the authors in [48], who demonstrated that the number of difficult words in the content had the highest impact on model accuracy. While Wikipedia encourages the use of plain English, difficult words are more commonly found in detailed, knowledge-intensive articles than in lower-quality ones. High-quality articles often require authors to provide in-depth knowledge on the topic, which frequently involves using technical language with a higher occurrence of difficult words. Regarding policy recommendations for writing Wikipedia articles, our findings suggest that using more difficult words does not necessarily guarantee higher-quality articles. However, if such words are needed to convey the topic effectively, authors should not hesitate to use them. Similarly, the activities on the talk pages, e.g., number of revisions, time elapsed between two revisions (i.e., the organisational attributes of Wikipedia) seem to be particularly important. Our belief is further strengthened by the following post-facto analysis of the results.

Success of our approach: We revisit a couple of articles and tried to compare the detected change points of quality switch predicted by our models with the ground truth. As noted in the previous sections, the talk pages contain heightened discussions among the editors before any quality change. Editors put forward their opinions, either converging to an agreement or landing at a conflict. We discuss two interesting cases here as depicted in the Figure 3.12. In the first case of *Catholic Church* article, the editor is found to oppose the current higher quality (AGA) and demoted the quality to B class, which is a prominent example of opinion conflict. In the other example, *Hurricane Hazel*, editors

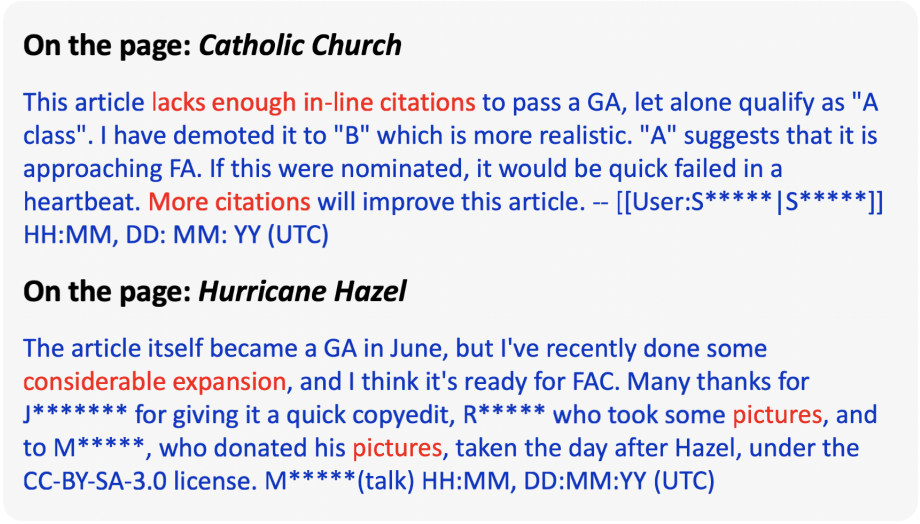


Figure 3.12: A snapshot of talk page conversations from a couple of randomly chosen pages that involve discussion for quality change of those pages.

had reached a consensus to promote the current quality of the article to the FA class. In both the cases, we include the timestamp mentioned in the talk pages as the change of quality and execute our methods. We observe that our proposed approach detect the change points accurately within the accepted margin of error M. Further, the detection of quality switch coincided with the actual change points, which we find is primarily triggered by an abrupt change in the content of the article.

Chapter 4

Knowledge equity across multilingual Wikipedia articles

In this chapter, we present a novel framework to transfer knowledge from a Wikipedia biography page of a high-resource language to that of a low-resource language. When the content in the high-resource language is also insufficient we make use of external resources, e.g., available books. Since these external resources do not always adhere to the NPOV standards we present a novel scheme to make them commensurate with Wikipedia guidelines before they could be integrated to the low-resource language article.

4.1 Understanding knowledge inequality

Wikipedia's decentralized structure and autonomous communities in various languages have established it as a global knowledge repository, often surpassing traditional and specialized linguistic resources. This influence is particularly significant in NLP research and development, where Wikipedia serves as a fundamental training data source for language models that emulate its style when generating texts or providing writing suggestions. Further, Wikipedia articles on the same topic in different languages are widely used

for various linguistic and translation tasks, such as statistical machine translation, text classification, and cross-lingual information retrieval. However, a significant knowledge gap exists across different language editions of Wikipedia [135], leading to an information divide among users of various language versions. For example, as of June 2024, the Hindi Wikipedia contains only 162,007 articles, despite Hindi being the third most spoken language globally and the most spoken language in India, with approximately 528.3 million speakers. In contrast, the English Wikipedia, the largest edition, has over 6.8 million articles, highlighting the disparity between high-resource and low-resource languages. A histogram illustrating the total number of Wikipedia articles across Indian languages vis-a-vis the English Wikipedia version as per Wikipedia statistics[1] is shown in Figure 4.1. Most of the information available online is in English and other high-resource languages, which is also reflected in Wikipedia. A significant portion of the world's population speaks low-resource languages, which are marginalized in the digital landscape. Despite the widespread availability and accessibility of internet resources, the scarcity of content in low-resource languages limits the ability of these communities to engage with online resources, preserve their culture, and access educational materials.

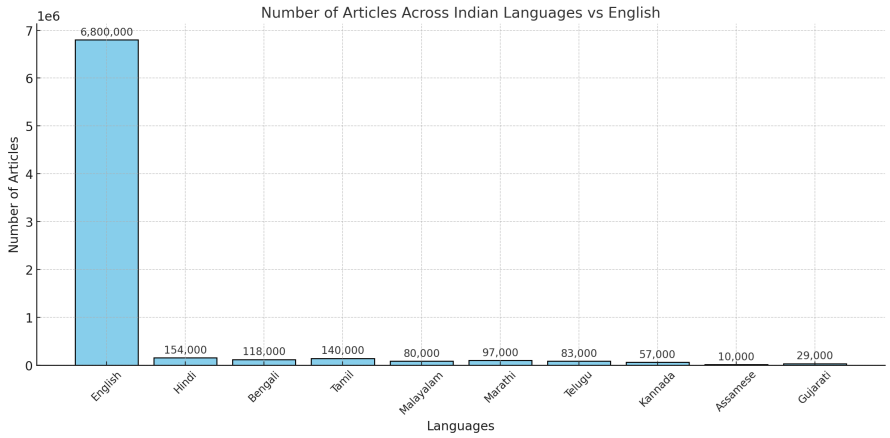


Figure 4.1: The histogram showing the number of Wikipedia articles across Indian languages compared to English language version.

Not only the lack in number of articles, previous studies [136, 166] have identified that the contents of Wikipedia articles on a particular topic in different language versions

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

can vary significantly. This content disparity can be attributed to several factors, such as the variability in the availability of information based on the cultural, historical, or regional relevance of that topic to the editors of a particular language. Second, the contributors to each language version may have different expertise levels, perspectives, or priorities, leading to variations in the coverage of an article in multiple languages. Identifying methodologies to address these disparities is a pivotal aspect of ensuring *knowledge equity*, a key concept introduced by the Wikimedia Foundation [163]. To assist contributors and address knowledge gaps in low-resource languages, it is possible to develop an automated cross-lingual framework that utilizes the reliable information available in English Wikipedia articles and further includes into Wikipedia articles written in low-resource languages both reliably and efficiently.

4.1.1 Our contribution

Our goal is to address the content disparity between high and low-resourced languages, in which English is assumed high-resourced language and Hindi is the low-resourced one. Despite both languages being among the top ten most widely spoken globally, a significant information gap exists in content coverage in Hindi biographies compared to their English counterparts. In light of this observation, our research is motivated by the key research questions.

- **RQ1:** Given a Wikipedia article in two languages, i.e., Hindi and English, how can one determine which language version lacks sufficient information and knowledge on that specific topic compared to the other?
- **RQ2:** How can one perform knowledge transfer from the more enriched language version to the less enriched one, usually high-resource to low-resource, automatically?

To address **RQ1**, we use article quality as an indicator to determine which language version contains more enriched information between the two. Not all Wikipedia language

versions, especially low-resource languages like Hindi in our study, have hierarchical quality scales to denote article quality. English Wikipedia is the most enriched and categorizes articles into seven quality classes as was discuss in the previous chapter: FA, A, GA, B, C, Start, and Stub (in descending order of quality). Therefore, we utilize language-agnostic article quality scores for each article as introduced by the authors in [52]. The underlying hypothesis is that lower quality scores indicate a need for content enrichment. In our dataset of Hindi and English Wikipedia articles, we select approximately 20,000 articles in which the Hindi version has a lower score compared to the English one.

To address **RQ2**, we propose a novel framework called WikiTransfer, which identifies content in English Wikipedia articles that can be efficiently transferred to the Hindi version. Specifically, we employ Large Language Models (LLMs) to facilitate accurate natural language conversion for transferring missing content from English to Hindi. WikiTransfer incorporates IndicTrans[2], a specialized model for multilingual translation in Indic languages, providing an efficient solution for cross-lingual content translation. Our pipeline, based on the quality of English articles, operates as follows.

Assume an article p is available in Hindi as H_p and in English as E_p.

- If E_p has been assessed as FA (i.e., of highest quality in English Wikipedia), we use the WikiTransfer framework solely for content transfer from E_p to H_p.
- If E_p belongs to any other quality class, we add two sub-modules: (a) external content augmentation and (b) external content neutralization, followed by the WikiTransfer framework.

As reliable source for external content, we collect publicly available books written in English for a subset of articles in our dataset. Next, we employ the retrieval component of the standard RAG (retrieval augmented generation) model to extract relevant information from these sources. External content often includes various writing biases, such as framing and epistemological biases, which can conflict with Wikipedia's NPOV policy.

[2]https://ai4bharat.iitm.ac.in/indic-trans2/

To address this, we leverage the in-context learning capabilities of off-the-shelf LLMs to generate neutral Wikipedia-style content. Finally, the neutralized text is machine-translated into Hindi with the help of WikiTransfer. This new information serves as the potential content for enhancing the articles in the Hindi version of Wikipedia biography. It is worth mentioning that our framework can generate content for individual sections without the involvement of any training. We evaluate the relevance of the newly generated Hindi content using heuristic-based metrics and human domain experts.

4.2 Dataset description

We employ a systematic approach to collect Wikipedia articles, which are available in both English and Hindi versions. We also anchor on the content of external resources for a subset of articles. In this work, we utilize a dataset comprising biographies, i.e., articles of Wikipedia category *people* sourced from Wikipedia. As the biography articles follow a predefined structure across multiple languages, thus enabling the development of WikiTransfer.

4.2.1 Collection of Wikipedia articles

Authors in [26] have published a large dataset of Wikipedia biographies in the ten most widely spoken languages in the world[3] (English, Mandarin Chinese, Hindi, Spanish, French, Standard Arabic, Bengali, Russian, Portuguese, and Indonesian). As we focus on the two languages – English and Hindi, we use this dataset and extract the biographies (i.e., article ID) with their corresponding Wikidata ID, which is available in the dataset in both language versions. These Wikidata IDs serve as unique identifiers, facilitating the identification of biography articles across different language versions. For example, the Wikidata ID for Serena Williams is *Q11459*. Using this ID, one can retrieve biography articles in all the listed languages in which the article is available. In our dataset, we

[3]https://www.visualcapitalist.com/100-most-spoken-languages/

amassed a total of 21,340 biography articles from the dataset that exist in both Hindi and English versions of Wikipedia. Next, we extract the wikitext (in both English and Hindi) of the current revision of these Wikipedia articles utilizing the MediaWiki API[4]. Later we pre-process the retrieved text and used the python package wikipedia[5] to extract the section headings for every article. Please note that we have not considered the sections– *See also*, *Notes*, *References*, *Further reading*, *External links* etc. and their corresponding sections in Hindi articles in our pipeline. In the plain text of the articles, the inline citations and hyperlinks (i.e., wikilinks to other articles) are discarded.

4.2.2 Collection of article quality

Utilizing the dataset [52], we collect quality scores for the Hindi and English versions of every article in our dataset. This dataset assigns a quality score between 0 and 1 to each article, regardless of language, with lower scores indicating poorer quality. We obtain a subset of 17,226 articles where the Hindi quality scores are lower than their English counterparts. This subset serves as our candidate set for further experiments. Since there exists no quality class hierarchy (unlike English Wikipedia) for Hindi Wikipedia articles, we leverage the language-agnostic quality scores.

Next, we extract the English quality categories, such as FA, A, GA, B, C, Start and Stub as predicted by ORES for each of the sampled articles. For the English articles in the FA category, we directly use their content to improve the lower-quality Hindi articles. For the English articles that are not in the FA category, our framework first enhances those articles using knowledge from external resources. Once these English articles are enriched, we use their improved content to augment into the corresponding Hindi versions. This approach ensures that the highest-quality information is transferred to enhance the less detailed Hindi articles.

[4]https://www.mediawiki.org/wiki/API:Get_the_contents_of_a_page
[5]https://pypi.org/project/wikipedia/

4.2.3 Collection of external resources

We utilize online digital libraries, particularly *Archive*[6], to source the biographical writings required for our enhancements. *Archive* provides a vast collection of scanned copies of historical books, making it an ideal resource for our purposes.

Automated search: We construct a search query consisting of the title of each biography article that refers to a link to the archive web page where the biographical writings of the specific article are stored. For example, for a particular biography, say P, the specific search query term – `https://archive.org/advancedsearch.php?q=P+AND+mediatype%3Atexts+AND+biography&fl[]=identifier&fl[]=title&rows=5&output=json` returns the corresponding biographical writings, say `bio`. We utilize the requests library and the HTTP method GET to extract the content of the web page. Further, we check the page content, if the keyword 'biography' exists, the response is considered valid. This approach ensures that our search results are more likely to include relevant documents from the Archive.

Manual verification: Due to the ambiguity in names and the nature of automated searches, many search results contain irrelevant or noisy information. To address this issue, we employ a post-graduate student who is a frequent Wikipedia user to verify the collected links manually. This step is crucial to ensure the quality and relevance of the biographical writings that we ultimately use. The manual verification process involves the following.

1. Reviewing search results for every query.
2. Filtering out noisy links irrelevant to the specific Wikipedia article or containing irrelevant information.

We then manually download the biographical writing in txt format and utilize the verified biographical writings to enrich Wikipedia biography articles. By integrating detailed and reliable information from these sources, we aim to significantly improve the quality of the English biographies followed by their Hindi versions.

[6] www.archive.org

Quality class (in English)	# of Articles	# of Biographical writings
FA	235	0
A	6	0
GA	485	13
B	1930	51
C	3428	38
Start	6625	0
Stub	4517	0

Table 4.1: Filtered dataset– articles categorized in quality classes and biographical writings extracted for the corresponding classes.

Thus, our meticulously curated dataset includes a sample of Wikipedia articles available in both English and Hindi versions, along with their respective quality scores. In addition, it contains a selective sample of biographical writings extracted from external resources. We tabulate the dataset statistics in the Table 4.1.

4.3 Proposed framework

We propose an end-to-end pipeline to transfer knowledge from English articles to their corresponding Hindi versions. The pipeline includes the framework WikiTransfer and additional modules for content transfer and external knowledge augmentation to enhance the Hindi version of articles.

4.3.1 WikiTransfer

Given an English and Hindi Wikipedia article, WikiTransfer works as follows.

1. First, it identifies section titles in the Hindi version that have semantic similarities with the titles in English.

2. Once the section mapping is completed, it checks whether the content of the mapped sections is semantically related. If the similarity between the section content in English and Hindi exceeds a certain threshold, the English content is machine-translated and appended to the existing Hindi section. If not, the translated content is suggested to be appended as a new section.

Section mapping: Since our framework operates at the individual section level, it is crucial to identify which section of an English page's content can be appended to a specific section of a Hindi page. To achieve this, we map the sections of English and Hindi pages. Mathematically, for an article p, let us assume the English page E_p and Hindi page H_p have m and n sections, respectively. To achieve the mapping between these sections, we first machine-translate the Hindi section titles to English using the IndicTrans model. Next, we compute the embeddings for each title and measure the cosine similarity between every pair of titles in the Hindi and English pages. For instance, for a Hindi title denoted as t_h, we compute the similarity with the embedding of m English section titles. The English section title with the highest similarity is then selected as the corresponding prospective content to be added to the Hindi section t_h. We use sentence transformers model[7] – all-MiniLM-L12-v2 to compute the embeddings of the section titles. We set the threshold to 0.44, which is the mean (μ) of the similarity scores; section pairs with similarity scores higher than the threshold are selected as the mapped sections.
Content matching: Even if the section titles are matched, we further analyze the content of the mapped sections to ensure coherence. Similar to section mapping, we compute the embeddings of the section content (Hindi and English) and find the cosine similarity. Section-content pairs with similarity scores exceeding the threshold are considered similar. We empirically select the threshold as $\mu + \sigma$ (mean: 0.89 and std dev: 0.06) of the similarity scores. For computing the embeddings of the section content, we use the multilingual model of sentence transformers multilingual e5-large[8].
Content augmentation: After establishing the section mapping and content mapping between English and Hindi versions, the subsequent step involves translating English

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2
[8]https://huggingface.co/intfloat/multilingual-e5-large

content into Hindi. For this purpose, we utilize the IndicTrans language model [68], which is pre-trained on a large corpus of Indic languages. Let us consider the English page E_p, which has e sentences in section t_e. These sentences are translated into e Hindi sentences, denoted as $Hindi(e)$. The corresponding mapped section t_h on the Hindi page H_p has an existing number of sentences denoted as h. The translated sentences $Hindi(e)$ are then appended to the existing h sentences in section t_h. However, before incorporating the machine-translated sentences, we implement a two-step filtering process. First, we discard translated sentences consisting of only one or two words, as they could potentially result from erroneous translation or lack of meaningful content. Next, for each sentence, say x in h, we identify the top three potential sentences among the translated sentences that are most semantically related to x. To achieve this, we compute the sentence embedding for both h sentences and $Hindi(e)$ sentences using the multilingual e5-base model of the sentence transformer. Likewise the previous mapping scheme, a cosine similarity score is calculated for each x with individual sentences in $Hindi(e)$ and we select the top three sentences among the sorted (in descending order) $Hindi(e)$ sentences that belong to the range of μ and $\mu + \sigma$ of the similarity scores. This way, we pick up three sentences that are somewhat dissimilar from the existing sentence x thus avoiding redundancy. This process prohibits overwhelming information from being incorporated into the Hindi section by discarding sentences that are nearly similar in the semantic space. We repeat the selection process for every sentence in h. Let us assume, if a sentence, say m in $Hindi(e)$ is selected in the top three list for both the sentences x and y of h, we consider the single occurrence of m in final appending phase of the sentences. Following this, we prepare a reduced set of $Hinid(e)$, denoted as $Hindi(e')$ which is added to the existing h sentences.

4.3.2 External content extraction

English articles that belong to quality classes other than FA might require additional information to enhance their quality. Therefore, we first gather information from external resources for such articles in our dataset and add it to the appropriate sections for further

processing using the WikiTransfer framework. Since our dataset focuses on individuals, we leverage personal narratives, such as biographical writings or autobiographies, as external references. For each English article, we source biographical writings available in English from the publicly accessible repository archive.org[9]. When multiple biographies are available for a given article, we select the most recently published version. To extract information from these external narratives aligned with the content of the articles, we employ the standard RAG method. Given a biographical Wikipedia page, we first select the corresponding personal narrative (autobiography or biography) as the external knowledge source and split the text (i.e., personal narrative) into several chunks of fixed length (we fix the length to 1000, 1200 characters) with a window of 200 characters allowed to overlap between the chunks using the function *RecusiveTextSplitter*[10]. Following this, we embed each of the chunks using `sentence-bert`[11] embeddings and store them in a vector database (we choose CHROMADB). Now, for the given English article E_p with m sections, we provide the content of the section t_i where $i \in 1, 2, .., m$ as query and external narratives W as the input to the *retriever* module of RAG pipeline. We use maximum marginal relevance (MMR) based search to retrieve top k chunks (we fix k to 3) relevant to the query. Out of these retrieved chunks, we utilize a suitable prompt (using Llama3(8B)-Instruct model), which identifies which prompt is the most relevant to the given section content. Here, it is worth mentioning that we do not use the text generator module of the RAG pipeline to generate the required content out of the retrieved chunk; rather, we use the POV rectifier module discussed below.

4.3.3 POV correction

Alongside Wikipedia's openness, a fundamental pillar of its success is its commitment to the NPOV policy, which ensures that facts should be presented fairly and impartially. According to this policy, Wikipedia prohibits sentences that contain perspective-specific or biased language, such as expressions of praise, criticism, or other sentiments that

[9] https://archive.org/

[10] https://github.com/langchain-ai/langchain

[11] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

reflect the editor's personal feelings. In Wikipedia's peer-reviewed environment, reviewers often use specific templates[12] to flag *POV* issues, guiding editors to address and resolve violations of Wikipedia's neutrality policy (*NPOV*). These issues are subsequently rectified by the respective editors. To accelerate this manual review process, researchers have developed various automated methods for detecting and mitigating biased sentences [6, 82, 87, 129, 130]. In this work, we adopt a straightforward approach using generative AI, specifically LLMs, to address *POV* issues effectively. Given that we are extracting content from biographies, which may include subjective language, there is a risk that the new content could violate Wikipedia's NPOV[13] standards. Therefore, to adhere to Wikipedia's NPOV policy, we identify and remove subjective biases, namely framing and epistemological biases [162] from individual sentences extracted from the biographies. If such biases exist we rephrase them in a way that adheres to the POV guideline. In this study, we attempt to leverage the power of LLMs to generate Wiki-style content. The most popular methods to use LLM in such downstream tasks are (1) supervised finetuning [94] and (2) in-context learning [167]. We perform our experiments with LLama-3(8B)-instruct model [5] for both these setups.

Supervised fine-tuning (SFT): For this setup, we finetune the LLama-3(8B)-instruct model using the WNC [158] and WikiBias [213] corpus. We use $\sim 2k$ data sample from WikiBias $Train_{manual}$ dataset and 10k biased and neutral sentence pairs randomly sampled from WNC corpus as training data to train the LLM.

In-context learning (ICL): For this setup, we use off-the-shelf instruction-tuned models namely LLama-3(8B) and LLama-3(70B). We use a generic prompt as shown below to generate a debiased sentence given a biased sentence. Specifically, we try (1) zero-shot (only the instruction) and (2) few-shot [155] (a few examples are used to describe the task to the model) prompting approach for the generation of Wiki-style neutral content.

[12] https://en.wikipedia.org/wiki/Template:POV

[13] https://tinyurl.com/cb7yv3tt

Prompt for generating Wiki-style sentence

For each query message, remove framing and epistemological biases, and do not add any extra content from your own knowledge.
Framing bias: subjective words or one-sided words, revealing the author's stance in a particular debate.
Epistemological bias: propositions that are either commonly agreed to be true or false and that are subtly presupposed, entailed, asserted or hedged in the text.
Here are some examples: . . .
Provide only the Output as: $< pad > output < /pad >$

Model	Methods	BLEU	METEOR	BERT
Llama3(8B)[SFT]	Zero-shot	0.27	0.5	0.94
	Few-shot	0.35	0.65	0.92
Llama3(8B)[ICL]	Zero-shot	0.25	0.6	0.93
	Few-shot	0.35	0.66	0.95
Llama3(70B)[ICL]	Zero-shot	0.24	0.57	0.93
	Few-shot	0.4	0.68	0.95

Table 4.2: Table showing evaluation score of Llama3 on test data in the two settings-- SFT and ICL. Among all the settings, Llama3(70B) in few-shot (5 shots) setup achieves the highest score for all three metrics.

Evaluation of the neutral content generator: Both of the configurations – SFT and ICL are evaluated on a sample of test data comprising 431 biased sentences and their neutral counterparts that exist in the WikiBias dataset. The WikiBias corpus includes a manually annotated subset known as WikiBias-Manual, which is further split into Train, Dev, and Test sets. For the evaluation, we utilized publicly available 431 biased sentences from the Dev set and their corresponding neutral counterparts. We evaluate the generated neutral content using three referential metrics – BLEU [152], METEOR [22],

and BERTScore [208]. The hypothesis behind employing these metrics is to assess the semantic similarity between the generated neutral sentences and the gold-standard human-annotated neutral sentences. The configuration achieving the highest score according to the above-mentioned metrics can be considered to produce neutral content equivalent to the ground-truth data, thereby aligning with Wikipedia's guidelines for maintaining a neutral point of view (NPOV). As evidenced in Table 4.2, ICL consistently outperforms SFT across all three reference-based metrics. Hence, we use the ICL few-shot setup (5 examples have been used in the prompt) and Llama3(70B) to generate neutral content for the extracted external book content, as mentioned in the above section.

Let us assume, for every section t_i in the English article E_p, we extract content from external resources and generate neutral content c_{nt} using the above-mentioned modules. If the section (in the English version) previously contained e sentences, after incorporating the neutral content c_{nt}, the combined content $e + c_{nt}$ of the section t_i is passed through WikiTransfer. This way, our pipeline facilitates knowledge transfer from the high-resource language (English) to the low-resource language (Hindi) in Wikipedia articles.

4.4 Evaluation setup and results

We evaluate the LLM-generated neutral Wiki-style NPOV content and the machine-translated Hindi content. For the former, we conduct human evaluation, and for the latter, we employ both automatic metrics and human evaluation.

4.4.1 Metric for automatic evaluation

To assess the relatedness of the newly generated content with the pre-existing Hindi text and the overall quality of the text, we utilize some standard metrics. The metric involves three information quality (IQ) components – *informativeness*, *readability*, *understandability* for evaluating the text content. Authors [178] in their work employed Google's

E-A-T framework for Wikipedia's IQ assessment, and three important factors of the framework are – expertise, authority, and trust. For the purpose of assessment of the machine-translated content, we valued expertise more than the other two factors. Here, expertise denotes how the content is exclusively knowledge-laden and informative and is composed by the contributors with relevant expertise on the subject matter. We adopt the following definition of components of expertise which was introduced in the above paper.

`informativeness` = 0.151 * `num_sentences` + 0.154 * `num_words` + 0.155 * `num_complex_words`
`readability` = 0.213 * `flesch_kincaid_grade` + 0.185 * `coleman_liau_index` + 0.26 * `complex_words_percentage` + 0.253 * `avg_syllables_per_word`
`understandability` = 0.393 * `gunning_fog_score` + 0.352 * `smog_index` + 0.181 * `automated_readability_index` + 0.344 * `avg_words_per_sentence`

However, in the above method, the simple *informativeness* metric does not take into account (a) how much new information has been added and (b) how appropriate the content is in continuing the existing section. Therefore, we experiment with an updated version of this metric – `calibrated informativeness` (CI) as proposed by the authors in [1]. CI is measured as follows:

- **`calibrated informativeness`** = Δ `informativeness` * `fraction_of_newly_added_words` * `continuation_score`

Here, the fraction of the newly added words determines how much new information has been added, and the continuation score determines how much the new content is appropriate in expanding the existing section content. Since, in the proposed metric, the difference of informativeness (Δ `informativeness`) between old and new content is considered, in our setting, we add the calculated CI to the informativeness of the old content and obtain `modified informativeness`. The new quality evaluation metric is tabulated as c_{new}^{cal} in Table 4.3, Table 4.4 and Table 4.8. Finally, the overall quality of the content of every section is calculated as follows, in which `informativeness`

is used in type c_{new} and `modified informativeness` is utilized in the type c_{new}^{cal}. The `modified informativeness` and `quality` metrics get reduced but are still far better compared to those of the old content.

quality = 0.255 * **(modified) informativeness** + 0.654 * **readability** + 0.557 * **understandability**

Because of the scarcity of resources, especially any unsupervised lexical method for assessing the quality of text in Hindi, we opted to measure quality in the English domain instead. We perform reverse translation for the newly generated Hindi section content along with the existing Hindi content into English and then evaluate the whole content utilizing the above-mentioned evaluation approach. For this purpose, we use the Python package *textstat*[14] for computing different statistical measures, such as `num_sentences`, `num_complex words`, etc. as well as measuring a few predefined metrics like `flesch_kinacid_grade`, `somg_index`, etc.

4.4.2 Results of automatic evaluation

FA quality articles: Following the WikiTransfer framework, let us assume the section t_e in the English version is mapped to the section t_h in the Hindi version and the newly generated content be c_{new} which is the combined content of existing Hindi text h and the newly machine-translated Hindi text $Hindi(e')$. Now, we perform reverse translation of both the new content c_{new} and the old content c_{old} (which is the set of existing Hindi h sentences). We then evaluate the two contents using the evaluation strategy stated above to assess their quality. We compute the metrics for individual sections and average over all the sections of the articles under consideration. The scores obtained using automatic evaluation are tabulated in Table 4.3. Overall we observe that the enhanced content is superior to the old content in terms of all the metrics.

[14] https://pypi.org/project/textstat/

Type	Informativeness	Readability	Understandability	Quality
c_{old}	43.86 (55.88)	4.62 (1.51)	16.42 (7.83)	23.35 (15.81)
c_{new}	87.57 (82.48)	4.87 (1.27)	17.94 (6.37)	35.50 (21.85)
c_{new}^{cal}	46.63 (55.68)	4.87 (1.27)	17.94 (6.37)	25.06 (14.85)

Table 4.3: Automatic evaluation: mean and (standard deviation) of the metric scores averaged over all the articles in our dataset that belong to *FA* quality class.

Given the high standard deviation observed in the informativeness metric for c_{old}, it is worthwhile to explore whether the improvements in content c_{new} compared to c_{old} are uniform across all the articles. To do this, we categorize the informativeness scores for c_{old} into three ranges based on their distribution and record the corresponding scores for the same sections in c_{new}. Table 4.4 displays the informativeness scores for both c_{old} and c_{new} across these three groups. It is clear that the informativeness has improved in c_{new} compared to c_{old} in each group, mirroring the results shown in Table 4.3.

Type	Group1 (0–50)	Group2 (50–100)	Group3 (100 and more)
c_{old}	18.6 (13.78)	71.00 (13.92)	177.18 (71.38)
c_{new}	61.79 (46.55)	108.22 (50.07)	235.18 (138.97)
c_{new}^{cal}	21.92 (14.05)	71.98 (14.13)	179.03 (73.57)

Table 4.4: Automatic evaluation: mean and (standard deviation) of the metric informativeness divided into ranges of scores for the articles that belong to *FA* quality class.

GA, B and C quality articles: For the articles that belong to the quality classes other than *FA*, we follow a similar evaluation strategy. However before we proceed for the automatic evaluation here we need check whether the external sentence have been effectively neutralized to obtain c_{nt}. To this purpose we perform a manual evaluation as follows.
Evaluation of LLM-generated NPOV content: We perform the human evaluation for 50 sentences randomly sampled from our dataset of content extracted from external resources. A few examples of biased sentences alongside their corresponding neutral counterpart sentences generated by the LLM are shown in Table 4.7. Evaluators are provided with each sentence pair, i.e., the original sentence as extracted from the external book corpus and the NPOV sentence as generated by the Llama3(70B) model.

Each evaluator assigns a score based on the scoring metric as mentioned in Table 4.5. Wikipedia's NPOV guidelines[15] state that any attribution or opinion must be supported by a citation to qualify it as a factual claim. This also applies to statements containing biased remarks about historically disadvantaged groups based on societal attributes. Typically, Wikipedia editors address such NPOV issues manually. In the absence of a manual process, automated models aim to produce neutral content that preserves the original meaning expressed by the author but removes any quantification or attribution. These statements can then be verified using external references or fact-checking tools. We employ two evaluators to manually assign the scores. We calculate the average score ($Score_{neu}$) obtained by each evaluator and report this in Table 4.6. Overall we observe that the neutralized content – c_{nt} obtained using our method is clearly suitable for augmentation to the destination Hindi page. The newly generated Hindi section

Original biased sentence: Blacks never listen to their parents.		
Score	**Rules**	**Example**
1	Complete bias removal	People do not always listen to their parents.
2	Complete bias removal + keeping the meaning (context) same	Some people never listen to their parents.
3	Complete bias removal + keeping the meaning (context) same + fluency	It is not uncommon for individuals to disregard parental advice.

Table 4.5: Scoring metric:This table presents the details of the scoring metrics used for annotation, along with examples based on a biased sentence. The original biased sentence is taken from the CrowS-Pairs dataset [141].

Human Evaluator	$Score_{neu}$
Evaluator 1	82.85%
Evaluator 2	77.14%

Table 4.6: Human evaluation on the generation of Wiki-style NPOV Hindi content through prompting.

content is c_{new}, which is the translated form of $e + c_{nt}$. We then compare this c_{new} with the old content c_{old} (both reverse translated) to assess the difference in terms of quality metrics. The result for automatic evaluation is mentioned in Table 4.8. Once again, we observe that c_{new} is superior to c_{old} across all the metrics.

[15]https://tinyurl.com/4sahkba4

Original Sentence	NPOV Sentence
He also took care to select exceptionally competent Vice-Chancellors like Sir J.C. Ghosh.	He also took care to select Vice-Chancellors like Sir J.C. Ghosh.
The places that Inayat frequented would not always be very interesting for his cousin	The places that Inayat frequented would not always be appealing to his cousin.
His marriage was of great importance in his life.	His marriage played a significant role in his life.
Sir William Hunter's career exhibited three well-defined phases.	Sir William Hunter's career exhibited three distinct phases.
He returned to Alma Mater as an assistant librarian but finally adopted a mercantile career, and died in 1885.	He returned to Alma Mater as an assistant librarian, but later pursued a career in business, and died in 1885.
This study has been for me a labour of love.	This study has been a significant undertaking for me.

Table 4.7: The table presents examples of original sentences extracted from external resources and their corresponding rectified NPOV sentences, i.e., neutral sentences generated by the POV correction module.

Type	Informativeness	Readability	Understandability	Quality
c_{old}	48.35 (56.94)	4.68 (0.89)	15.79 (3.73)	24.19 (14.98)
c_{new}	113.67 (67.29)	4.93 (0.69)	17.25 (2.93)	41.82 (17.36)
c_{new}^{cal}	50.83 (56.67)	4.93 (0.69)	17.25 (2.93)	25.97 (15.20)

Table 4.8: Automatic evaluation: metric scores averaged over all the sections in our dataset of English articles – GA, B and C quality classes.

4.4.3 Results of manual evaluation

We assess the newly generated content by our framework for both the settings – (a) augmentation of new content from *FA* articles only and (b) augmentation of new content from the articles other than *FA* along with the external content. The evaluation focuses on three key qualitative metrics – (a) informativeness, (b) readability, and (c) coherence – each rated on a scale from 1 to 3. Seven Hindi-speaking evaluators are tasked with this assessment. The evaluation scheme is outlined as follows.

- **Informativeness**: This indicates the ability of a piece of text to provide useful information and comprehensive content. To evaluate the informativeness, the evaluators have to compare the original Hindi content c_{old} consisting of existing Hindi sentences (h) and the newly generated Hindi content c_{new} of the mapped section. If c_{new} provides similar information without any additions, the informativeness

score should be 1. However, if c_{new} includes additional information not present in c_{old}, evaluators will assign a score of 2 for moderate additions or 3 for a higher level of new information added.

- **Readability**: This metric measures the effort required by the reader to read and understand a piece of information. If the vocabulary and sentence structure in the text are complex, the difficulty of reading increases. Our human evaluators assess the readability of c_{new} on a scale from 1 to 3, with the understanding that lower readability corresponds to a lower score.

- **Coherence**: This metric represents the logical flow between sentences in a text, ensuring that they naturally follow one another to form meaningful content. Similar to the readability metric, the coherence of c_{new} is evaluated on a scale from 1 to 3, with higher scores indicating greater coherence.

We randomly sample 35 sections from our dataset to evaluate the content generated by our framework. Among these, 10 sections are from the FA category, while the remaining 25 sections are chosen from non-FA quality categories: 5 from GA, 10 from B, and 10 from C. The human assessments on the above-mentioned three metrics are tabulated in Table 4.9. The average informativeness, readability, and coherence across the seven human judges, respectively, are – (FA: 2.67, non-FA: 2.68), (FA: 2.46, non-FA: 2.50), and (FA: 2.34, non-FA: 2.32). Thus, we observe that for all the metrics and for both FA and non-FA categories, the average judgments are always 2.3+, indicating the newly generated content is reasonably good in terms of the three metrics, especially informativeness and readability. We find significant improvement in informativeness for both FA and non-FA groups, suggesting effective addition of relevant knowledge to existing sections. Given our multi-label evaluation scheme and involving multiple annotators, we compute inter-annotator agreement using Fleiss' Kappa method [66]. We obtain κ values of 0.61, 0.53, and 0.54 for informativeness, readability, and coherence, respectively, indicating a moderate to substantial level of agreement among annotators in assessing the generated content. We present the Hindi content from an existing Wikipedia article alongside the content generated by our framework for two sample sections: one from

Type	Informativeness		Readability		Coherence	
	FA	non-FA	FA	non-FA	FA	non-FA
Evaluator 1	2.9	2.77	2.8	2.5	2.5	2.43
Evaluator 2	2.2	2.3	2.1	2.5	1.8	2.17
Evaluator 3	3	2.93	2.4	2.33	2.3	2.0
Evaluator 4	2.6	2.7	2.6	2.7	2.6	2.53
Evaluator 5	2.4	2.37	1.9	2.17	2.0	2.03
Evaluator 6	3	3	2.7	2.67	2.8	2.8
Evaluator 7	2.6	2.67	2.7	2.6	2.4	2.3

Table 4.9: Human evaluation on the generated machine-translated Hindi content based on three metrics – informativeness, readability, coherence.

a featured article (FA) and the other from a C-class article. The Hindi output for the FA article is generated using WikiTransfer, and both the Hindi content and its English translation are displayed in Figure 4.2. For the C-class article, the new Hindi content is created by first pooling in text from external resources using the RAG method, followed by the NPOV correction and the WikiTransfer framework. The corresponding Hindi output and its English translation for this sample section are presented in Figure 4.3.

4.5 Summary

To address the content disparity between high-resource and low-resource language editions of Wikipedia, we proposed a novel solution focused on individual sections of Wikipedia articles. Our framework first incorporates existing human-edited knowledge from English articles, particularly featured articles (FA) that contain verifiable facts, neutral tone, and expert assessment, making them ideal for augmentation into low-resource languages. Further, our framework extracts relevant content from external sources in high-resource languages like English for knowledge transfer, rather than relying directly on generative AI systems. This extracted content is adapted to Wikipedia's style and NPOV policy using the in-context learning capabilities of LLMs. Finally, the combined knowledge (existing and newly extracted content) in the high-resource language is machine-translated into the target low-resource language. This translated content is then ready

Existing Hindi	Newly generated Hindi
फ्रीडा पिंटो मुंबई में सिल्विया पिंटो, जो सेंट जॉन यूनिवर्सल हाई स्कूल (गुडगाँव) में प्रधानाचार्य हैं और फ्रेडरिक पिंटो, जो की बड़ौदा बैंक के एक वरिष्ठ शाखा प्रबंधक हैं उनके घर पैदा हुई। फ्रीडा पिंटो के पिता नीरुड़े से और उनकी माता डेरेबैल से हैं, यह दोनों कस्बे मेंग्लोर में हैं। उनका परिवार मेंग्लोरीयन कैथोलिक मूल का है, और एक इंटरव्यू में, पिंटो ने कहा है कि वह पूरी तरह से शुद्ध भारतीय हैं, लेकिन उनका परिवार कैथोलिक है। उनकी बड़ी बहन शेरोन पिंटो NDTV समाचार चैनल पर एक सहायक निर्माता है। पिंटो ने मलाड में सेंट जोसेफ स्कूल के कर्मेल में अभ्यास किया और अंग्रेजी साहित्य में कला स्नातक (BA) की उपाधि सेंट जेवियर्स कॉलेज, मुंबई से ली। वर्तमान समय में वह मुंबई के एक उपनगर, मलाड के ऑरलेम इलाके में रहती है। वह भारतीय शास्त्रीय नृत्य के विभिन्न रूपों के साथ साथ साल्सा में भी प्रशिक्षित हैं। वह अपने पूर्व प्रचारक और आठ साल के प्रेमी रोहन अन्ताओ के साथ संलग्न थी, लेकिन 2009 की शुरुआत में स्टार बनने के बाद उन्होंने यह सगाई तोड़ दी। आजकल वह अपने 'स्लमडॉग मिलियनेर' फिल्म के सह अभिनेता, देव पटेल के साथ डेटिंग कर रहीं है, जो की उनसे उम्र में छह साल छोटे है। वह पीपल People पत्रिका की सबसे खूबसूरत लोगों की सूची में और दुनिया की बेहतरीन सजने वाली महिलाओं की सूची में भी शामिल हो चुकी हैं।	फ्रीडा पिंटो मुंबई में सिल्विया पिंटो, जो सेंट जॉन यूनिवर्सल हाई स्कूल (गुडगाँव) में प्रधानाचार्य हैं ओर फ्रेड्रिक पिंटो, जो की बड़ौदा बैंक के एक वरिष्ठ शाखा प्रबंधक हैं उनके घर पैदा हुई। फ्रीडा पिंटो के पिता नीरुड़े से और उनकी माता डेरेबैल से हैं, यह दोनों कस्बे मेंग्लोर में हैं। उनका परिवार मेंग्लोरीयन कैथोलिक मूल का है, ओर एक इंटरव्यू में, पिंटो ने कहा है कि वह पूरी तरह से शुद्ध भारतीय हैं, लेकिन उनका परिवार कैथोलिक है। उनकी बड़ी बहन शेरोन पिंटो NDTV समाचार चैनल पर एक सहायक निर्माता है। पिंटो ने मलाड में सेंट जोसेफ स्कूल के कर्मेल में अभ्यास किया और अंग्रेजी साहित्य में कला स्नातक (BA) की उपाधि सेंट जेवियर्स कॉलेज, मुंबई से ली। वर्तमान समय में वह मुंबई के एक उपनगर, मलाड के ऑरलेम इलाके में रहती है। वह भारतीय शास्त्रीय नृत्य के विभिन्न रूपों के साथ साथ साल्सा में भी प्रशिक्षित हैं। वह अपने पूर्व प्रचारक और आठ साल के प्रेमी रोहन अन्ताओ के साथ संलग्न थी, लेकिन 2009 की शुरुआत में स्टार बनने के बाद उन्होंने यह सगाई तोड़ दी। आजकल वह अपने 'स्लमडॉग मिलियनेर' फिल्म के सह अभिनेता, देव पटेल के साथ डेटिंग कर रहीं है, जो की उनसे उम्र में छह साल छोटे है। वह पीपल People पत्रिका की सबसे खूबसूरत लोगों की सूची में और दुनिया की बेहतरीन सजने वाली महिलाओं की सूची में भी शामिल हो चुकी हैं।. के लिए काम करती हैं।पिंटो की परवरिश उत्तरी मुंबई के मलाड उपनगर में हुई थी।वह पहली बार एक अभिनेत्री बनना चाहती थी जब वह पाँच साल की थी, अक्सर अपने बचपन के दौरान टेलीविजन अभिनेताओं के कपड़े पहनती थी और उनकी नकल करती थी।बाद में उन्होंने 1994 की मिस यूनिवर्स प्रतियोगिता में सुष्मिता सेन की जीत से प्रेरित होने को याद करते हुए कहा कि देश को वास्तव में उन पर गर्व था, और मुझे लगा, एक दिन, मैं भी ऐसा ही करना चाहती हूं।पिंटो ने सेंट के कार्मेल में भाग लिया।मलाड, उत्तरी मुंबई में जोसेफ स्कूल और फिर सेंट में अध्ययन किया।जेवियर्स कॉलेज, मुंबई, फोर्ट, दक्षिण मुंबई।उनका प्रमुख अंग्रेजी साहित्य में था, जिसमें मनोविज्ञान और अर्थशास्त्र में नाबालिग थे।कॉलेज में, उन्होंने शौकिया रंगमंच में भाग लिया, लेकिन 2005 में स्नातक होने तक अभिनय ओर मॉडलिंग के कार्यों को अस्वीकार कर दिया। कम उम्र से ही अभिनय में उनकी रुचि के बावजूद, पिंटो ने यह तय नहीं किया था कि कॉलेज में मॉन्स्टर (2003) देखने तक कौन सा करियर लेना है।

(a)

Existing Hindi (English translation)	Newly generated Hindi (English translation)
Freida Pinto was born in Mumbai to Sylvia Pinto, who is the principal at St. John's Universal High School (Gurgaon) and Frederick Pinto, a senior branch manager at Baroda Bank. Freida Pinto's father is from Neerude and her mother is from Derebail, both towns in Mangalore. His family is of Mangalorean Catholic origin, and in an interview, Pinto has stated that he is a fully pure Indian, but his family is Catholic. Her elder sister Sharon Pinto is an assistant producer on the NDTV news channel. Pinto studied at St. Joseph's School Carmel in Malad and graduated with a Bachelor of Arts (BA) in English literature from St. Xavier's College, Mumbai. Currently, she lives in the Orlem area of Malad, a suburb of Mumbai. She is trained in various forms of Indian classical dance as well as Salsa. She was engaged to Rohan Antao, her former publicist and boyfriend of eight years, but broke off the engagement after she became a star in early 2009. She is currently dating her Slumdog Millionaire co-star, Dev Patel, who is six years her junior. She has also appeared on People magazine's list of the most beautiful people and on the list of the world's best-dressed women.	Freida Pinto was born in Mumbai to Sylvia Pinto, who is the principal at St. John's Universal High School (Gurgaon) and Frederick Pinto, a senior branch manager at Baroda Bank. Freida Pinto's father is from Neerude and her mother is from Derebail, both towns in Mangalore. His family is of Mangalorean Catholic origin, and in an interview, Pinto has stated that he is a fully pure Indian, but his family is Catholic. Her elder sister Sharon Pinto is an assistant producer on the NDTV news channel. Pinto studied at St. Joseph's School Carmel in Malad and graduated with a Bachelor of Arts (BA) in English literature from St. Xavier's College, Mumbai. Currently, she lives in the Orlem area of Malad, a suburb of Mumbai. She is trained in various forms of Indian classical dance as well as Salsa. She was engaged to Rohan Antao, her former publicist and boyfriend of eight years, but broke off the engagement after she became a star in early 2009. She is currently dating her Slumdog Millionaire co-star, Dev Patel, who is six years her junior. She has also appeared on People magazine's list of the most beautiful people and on the list of the world's best-dressed women. working for. Pinto grew up in the Malad suburb of North Mumbai. She first wanted to be an actress when she was five years old, often dressing up and mimicking television actors during her childhood. She later recalled being inspired by Sushmita Sen's victory at the 1994 Miss Universe pageant, saying that the country was \"really proud of her, and I thought, one day, I want to do the same.\"Pinto attended the Carmel of Sts. Joseph School in Malad, North Mumbai, and then studied at St. Xavier's College, Mumbai, Fort, South Mumbai. Her major was in English literature, with minors in psychology and economics. In college, she participated in amateur theatre but declined acting and modelling assignments until graduating in 2005. Despite his interest in acting from an early age, Pinto did not decide which career to pursue until watching Monster (2003) in college.

(b) English translation of Hindi content

Figure 4.2: An example of existing and WIKITRANSFER generated new content– a sample section that belongs to FA quality– (a) Hindi version, (b) English version

Existing Hindi	Newly generated Hindi
1863 में गुरु विरजानंद के पास अध्ययन पूर्ण होने के बाद लगभग बीस वर्षों के कार्यकाल में दयानंद सरस्वती की हत्या व अपमान के लगभग 44 प्रयास हुए। जिसमें से 17 बार विभिन्न माध्यमों से विष देकर प्राण हरण के प्रयास हुए।	1863 में गुरु विरजानंद के पास अध्ययन पूर्ण होने के बाद लगभग बीस वर्षों के कार्यकाल में दयानंद सरस्वती की हत्या व अपमान के लगभग 44 प्रयास हुए। जिसमें से 17 बार विभिन्न माध्यमों से विष देकर प्राण हरण के प्रयास हुए।. दयानंद को अपनी जान से मारने के कई असफल प्रयासों का सामना करना पड़ा।उनके समर्थकों के अनुसार, उन्हें कुछ मौकों पर जहर दिया गया था, लेकिन हठ योग के नियमित अभ्यास के कारण वे इस तरह के सभी प्रयासों से बच गए।एक कहानी बताती है कि हमलावरों ने एक बार उसे नदी में डुबोने का प्रयास किया था, लेकिन दयानंद ने हमलावरों को नदी में घसीटा, हालांकि उसने डूबने से पहले उन्हें छोड़ दिया।एक अन्य विवरण में दावा किया गया है कि उन पर मुसलमानों ने हमला किया था जो गंगा पर ध्यान देते समय इस्लाम की आलोचना से नाराज थे।उन्होंने उसे पानी में फेंक दिया लेकिन दावा किया जाता है कि उसने खुद को बचाया क्योंकि उसके प्राणायाम अभ्यास ने उसे हमलावरों के जाने तक पानी के नीचे रहने दिया।\"मैं सत्य का शिष्य हूँ।मैं हमेशा सच बोलूंगा।अगर ऐसा करने में मैंने किसी को चोट पहुंचाई है तो गलती मेरी नहीं है।मैं सच को कभी नहीं छोड़ूंगा।उन्होंने कहा, \"वह अपने शब्दों पर खरे उतरे।उनके उपदेशों के कारण, कई लोग दयानंद के अनुयायी बन गए।उनमें से कई लोगों ने उनके शिष्य बनने का फैसला किया।न केवल हिंदुओं के बीच बल्कि मुसलमानों और ईसाइयों के बीच भी उनके समर्थक थे।उनमें रेवरेंड स्कॉट और सैयद अहमद खान उल्लेखनीय हैं।हालांकि, कई लोगों ने उनका विरोध किया।वे उसे अपने हितों के लिए खतरा मानते थे।वे उसे धमकी देते थे; उसके भोजन और दूध को जहर देते थे और अक्सर हिंसा के साथ उसका सामना करते थे।उन्होंने शांति से जवाब दिया और उन लोगों के प्रति दया दिखाई जो उन्हें नुकसान पहुँचाने की कोशिश कर रहे थे।

(a)

Existing Hindi (English translation)	Newly generated Hindi (English translation)
In 1863, after completing his studies with Guru Virajanand, there were about 44 attempts on Dayanand Saraswati's life over a period of about twenty years. Out of which, 17 attempts were made to kill by poisoning through various means.	In 1863, after completing his studies with Guru Virajanand, there were about 44 attempts on Dayanand Saraswati's life over a period of about twenty years. Out of which, 17 attempts were made to kill by poisoning through various means. Dayanand faced several unsuccessful attempts to kill himself. According to his supporters, he was poisoned on a few occasions, but he escaped all such attempts due to his regular practice of Hatha yoga. One story states that the attackers had once tried to drown her in the river, but Dayanand dragged the attackers into the river, though he abandoned them before drowning. Another account claims that he was attacked by Muslims who were outraged by criticism of Islam while meditating on the Ganges. They threw him into the water but it is claimed that he saved himself as his pranayama practice left him underwater until the attackers left.\"I am a disciple of truth. I will always tell the truth. If I have hurt someone in doing so, it is not my fault. I will never give up on the truth.\"He lived up to his words. Due to his teachings, many became followers of Dayanand. Many of them decided to become his disciples. He had supporters not only among Hindus but also among Muslims and Christians. Notable among them are Reverend Scott and Syed Ahmad Khan. However, many people opposed it. They considered it a threat to their interests. They threatened her; poisoned her food and milk and often confronted her with violence. They responded calmly and showed kindness to those who were trying to harm them.

(b) English translation of Hindi content

Figure 4.3: An example of existing and our framework generated new content– a sample section that belongs to C quality– (a) Hindi version, (b) English version

for integration into Wikipedia articles in the target language. Our lightweight framework produces content that is substantially superior to the existing content for low-resource language articles, as per the evaluation based on automated metrics and human assessments.

Chapter 5

Assessment of bias and fairness in Wikidata

In this chapter, we investigate the social biases embedded in knowledge graphs, specifically in Wikidata. The first problem attempts to identify this in the form of data and algorithmic biases present in Wikidata. The second problem investigates the influence of social biases on downstream applications, such as link prediction. In both problems, we examine the variations resulting from two key design choices: (a) the choice of geo-social data and (b) the embedding learning algorithms, along with their universal characteristics.

5.1 Societal biases in Wikidata as data bias and algorithm bias

With the rapid expansion of content available on the web, *Knowledge Graph* (KG) is being widely used by tech giants to assimilate factual information in a structured format that can be used in many industrial applications. Academic research on KGs has been gaining significant interest for exploring knowledge in the light of representation of large-

scale graph-structure data, reasoning, and application in a variety of automated tasks - recommendation [192], chatbot question answering [86], language modelling [156, 211] etc. In almost every downstream task of machine learning, knowledge extracted from knowledge graphs is assumed as the gold standard that establishes the correctness of the typical system. Like DBpedia[1], Google Knowledge Graph[2], Wikidata[3] is an open collaborative knowledge base that are created to serve as the hub of structured linked data to all Wikimedia projects as well as to external digital tools and bots. Over the years, based on the efforts of human editors and automated software, Wikidata has become a giant knowledge base of over 99 million entities in multiple languages[4]. However, while knowledge graphs are evolving continually, the user-generated content of knowledge graphs mediates societal disparities in many downstream applications. For example, authors in [205] have found that less than 22% of Wikidata items represent people who are women and thus present a severe gender disparity in its content.

Bias in KGs: Bias refers to the systematic and unfair treatment to certain individuals, groups, or perspectives that arises in the context of human-computer interactions, particularly in systems that mediate, process, or analyze social behavior. Bias can manifest in various forms, including algorithmic bias, data bias, cognitive bias etc., significantly affecting system fairness and societal outcomes. A system engages in unfair discrimination if it denies an individual or group access to opportunities or benefits, or assigns undesirable outcomes, based on unreasonable or inappropriate criteria [147]. Detecting bias starts with the data set. Data bias emerges when the input data is incomplete, imbalanced, or reflective of existing societal prejudices [24, 34, 39, 41, 132]. Further, algorithmic bias occurs when computational models make decisions that disadvantage specific groups due to inherent patterns in training data or flawed algorithms [146]. Algorithmic design choices, such as the selection of optimization functions, regularization techniques, decisions to apply regression models to the entire dataset versus specific subgroups, and the use of statistically biased estimators in algorithms, can all influence and contribute

[1] https://www.dbpedia.org/resources/knowledge-graphs/
[2] https://developers.google.com/knowledge-graph
[3] https://www.wikidata.org/wiki/Wikidata:Main_Page
[4] https://www.wikidata.org/wiki/Wikidata:Statistics

to biased decisions, ultimately affecting the outcomes of the algorithms [51]. Further, one needs to note that technical systems are not neutral; they reflect the perspectives and values embedded within the society that designs and uses them [58]. Knowledge Graphs (KGs) are considered factual if they accurately reflect the real world. For instance, a KG containing only U.S. presidents truthfully represents historical data. However, such a snapshot could lead to biased inferences, like predicting that only men can become presidents, despite U.S. laws permitting candidates of any gender. This reliance on historical data risks reinforcing societal inequalities and undermining social fairness by focusing on "what is" rather than "what should be." Knowledge graph embeddings, which summarize statistical and distributional patterns, further encode these biases. For example, if a KG predominantly includes male presidents, embeddings may infer a strong association between the gender 'male' and the role 'president,' perpetuating gender bias in predictive models. In fact, a wide range of societal and human biases can be attributed to the curation of KGs. The entities and relationships in a typical KG are accumulated in a (semi) automatic way [55, 91], which may result in gathering biased knowledge from the open text corpus of the web. Furthermore, algorithms used to sample, aggregate, and process knowledge can incorporate biases into KGs. In handshake with the proliferation in embedding learning methods [47, 93], recent works establish the anecdotal presence of societal biases in KG data and how they are being mirrored by state-of-the-art KG embedding algorithms [33, 65]. Biases encoded in KGs and knowledge graph embeddings (KGEs) have a negative effect on society as well as the underlying automation systems that leverage the knowledge extracted from KGs in building the system. As knowledge graph embeddings are increasingly used as inputs to large language models, their inherent biases can influence downstream tasks, much like biases in word embeddings. To tackle this issue, researchers come up with coherent frameworks of bias measurement and debiasing them further [17]. Previous works [33, 65] on identifying biases are mostly focused on specific KGs or typical KG embedding learning algorithms, but their further comparison in the context of sensitive attributes and differences of societal constraints is missing in the literature. There could be such scenarios in which biases incurred from a sensitive attribute can vary across socio-economic, socio-cultural, and geographical boundaries fixing the knowledge graph upfront. For instance, we have found evidence

in our experiment that the occupation "activist" is a female-dominated occupation in the Middle Eastern geographies while it is a gender-neutral occupation in the Western world considering the fact that *gender* is a sensitive attribute. Building on works such as [60, 133], which examine social biases in large language models through the lens of geographical contexts, it would be compelling to investigate how these biases manifest and vary within knowledge graphs when analyzed across geographical boundaries.

Our contributions: In this work, we hypothesize that the choice of geo-social data, i.e., knowledge graphs pertaining to different geographies and KG embedding representation algorithms has a significant influence on the behavior of bias measurement in KGs and may lead to notable variability of biases depending on design choices. To demonstrate the variance of societal biases that exist in KGs, we introduce an empirical data-driven analysis on Wikidata. Among a broad set of sensitive attributes, gender is selected for our analysis since it leads to biases. The stereotypical observable that we consider here is professions or the dominated occupations by a specific gender. We assume that the base dataset in our experiment itself contains biases resulting from cultural differences (aka *data bias*). Note that such biases are inherent and cannot be, therefore, controlled in a crowd-sourced system. In this context, we have the following objectives-

- As the first objective, our point of interest is to check whether *algorithmic biases* creep in when we build embedding from this base dataset. The hypothesis is that if they do, then the data bias ought to be altered.
- The second objective of this work is to audit the existence of gender bias (based on gender division - male and female) across two important orthogonal axes - (i) choice of training data modality and (ii) embedding learning algorithm.

For this purpose, we collect a large number of Wikidata entities and relations from 13 different geographies around the globe - Arabia, Australia, Argentina, Brazil, France, Germany, India, Japan, Kenya, Russia, South Africa, United Kingdom, and the United States of America - ensuring representation from across the different continents. We conduct our experiments using two important knowledge graph embedding algorithms -

TRANSE [30] and COMPLEX [182]. Our hypothesis is confirmed by our analysis; below, we list the contributions of this work.

- We observe that the inherent data bias present in the base dataset is indeed revised by the embedding learning algorithm demonstrating the presence of algorithmic biases.
- With respect to a specific bias measurement metric, the ranked lists of female/male-dominated occupations are highly sensitive to both the embedding algorithm – TRANSE and COMPLEX and the underlying data geography.

5.1.1 Background

Formally, a knowledge graph is structured as a graph abstraction with nodes representing entities from a given domain and edges corresponding to relations between the entities. Ideally, facts in the knowledge graph are described as labeled triple formats, such as $< h, r, t >$ where h and t are the head and tail entities, respectively, and r denotes the relation between h and t. For example, the knowledge triple <*Joe Biden, born_in, US*> expresses the fact that *Joe Biden* was born in *US*. Among the existent large number of knowledge graph embedding algorithms aimed at learning dense representations of the knowledge graphs, we use two popular KG embedding algorithms, namely TRANSE and COMPLEX in our experiment for simplicity and scalability of use.

1. TRANSE [30]: The most elementary approach in KG embedding is the use of *translational models* that assume a geometric perspective in which relation embeddings translate subject entity and relation to object entity in the low-dimensional space. The loss function f_{TRANSE} is defined as the L_1 or L_2 norm between the embedding of the tail and the embedding of the head plus the embedding of the relation as follows - $f_{\text{TRANSE}} = -||h + r - t||_n$.
2. COMPLEX [182]: This algorithm relies on *tensor decomposition method*, and each entity and relation is assumed as a complex vector (i.e., a vector containing

complex numbers) of dimension d. The loss function of COMPLEX is denoted as follows - $f_{\text{COMPLEX}} = Re(< r, h, \bar{t} >)$. Here, Re denotes the real vector component of the embedding generated in the complex space.

Measuring data bias in KG

To understand how biases are encoded in the knowledge triples of Wikidata, we measure the data bias in our dataset of 13 geographies collected from Wikidata as proposed by Bourli [33] et al. We briefly describe the metric below.

The metric: Let us assume a KG dataset contains two types of triples primarily– <*human_entity, has_gender, gender*> and <*human_entity, has_occupation, occupation*>. First, the metric computes the bias score θ for every occupation in the dataset. Formally, let us assume that for an occupation o, the number of male (female) entities that have occupation o be M_o (F_o) and M (F) be the total number of male (female) entities that exist in the dataset. Now, given the occupation o, let $Pr(O = o|G = m)$ and $Pr(O = o|G = f)$ be the respective probabilities that male and female entities have the occupation o, where O, G, m, f denote occupation, gender and its binary attributes - male and female respectively. For a given dataset, the probabilities and bias score are computed as follows – (i) $Pr(O = o|G = m) = |M_o|/|M|$, (ii) $Pr(O = o|G = f) = |F_o|/|F|$ and (iii) $Pr(O = o|G = m) - Pr(O = o|G = f) = \theta$.

For a specific dataset, a threshold t is estimated, t being a positive value close to 0. For the occupation o, if the bias score $\theta > t$, the occupation is assumed to be male-biased, otherwise, if $\theta < -t$ then it is female-biased. Further, the occupations that have a bias score within the range $[-t, t]$ are treated as neutral occupations. The threshold t is selected from the distribution[5] of neutral occupations for different values of t.

[5]The distribution is generated by plotting the count of neutral occupations for different values of t, and the optimal t is selected as the value for which there is a sharp increase in the count of the neutral occupations.

Measuring biases in KG embedding

Similar to the existence of social biases in word embeddings [29], the biases in trained knowledge graph embeddings are prone to discriminative notions; for example, the occupation 'banker' is more strongly tied to male than female entities. It results in many downstream applications skewed toward a particular section (e.g., based on gender) of society. Our objective is to investigate the influence of gender on the embedding method's outcome in predicting whether an occupation is male or female-biased, or gender-neutral. In order to measure gender biases encoded in graph embedding of Wikidata, we followed one of the very popular bias measurement approaches proposed by Fisher et al. [65], which is briefly described below.

Bias metric by Fisher et al. [65]: According to this metric, the first step is to update the initial embedding of a person in the knowledge graph to increase the representation of the male component in the person's embedding. This embedding could be a pretrained one or a representation uniformly sampled from an arbitrary distribution. The updation of embedding is achieved by providing the model with two batches of triples, (e_j, r_g, e_a) and (e_j, r_g, e_b), where e_j is the embedding of the person j, r_g is the embedding of the sensitive attribute (i.e., gender) and e_a, e_b denotes the embedding of two primary values of the attribute gender, male and female respectively. The score function is denoted by $g(.)$, which takes the embeddings of a triple as input and outputs a score, denoting how likely this triple is to be correct. Next, we differentiate the score m with respect to the embedding of person j, e_j, and take a step in the stochastic gradient descent algorithm with a learning rate α in order to maximize the score function as described in equation 5.1 and equation 5.2.

$$m(\theta) = g(e_j, r_g, e_a) - g(e_j, r_g, e_b) \tag{5.1}$$

$$e'_j = e_j + \alpha \frac{\delta m(\theta)}{\delta e_j} \tag{5.2}$$

The second step is to calculate the difference of the scores given by the scoring function on the following triples (e'_j, r_p, e_p) and (e_j, r_p, e_p) where e'_j is the updated embedding at the end of the updation process, r_p is the embedding of the relation *has_occupation* and e_p denotes the embedding of an occupation. Thus it assigns a bias score b_p (i.e.,

the difference) to the occupation p. The above-mentioned steps are reiterated for all the human entities in an underlying knowledge graph comprised of two types of entities in general–<*human_entity, has_gender, gender*> and <*human_entity, has_occupation, occupation*>. Finally, the occupations that exist in the knowledge graph are ranked in descending order of their bias scores.

Formally, let us assume that we have two occupations p_1 and p_2 in our sample knowledge graph, and the number of human entities is N. We calculate the bias scores b_{p_1} and b_{p_2} for the occupations p_1 and p_2 respectively following the steps as mentioned in the above bias measurement metric. In both cases, the bias score is computed over the set of all human entities N in the knowledge graph. Now, we rank the bias scores and let us assume $b_{p_1} > b_{p_2}$. Hence, p_1 is ranked as a higher male-biased occupation than p_2. In other words, more the bias score, an occupation is closer to the gender line (i.e., the male part e_a of the sensitive attribute gender than the female part e_b in equation 5.1). The ranking of occupations in terms of descending order of bias score shows the likelihood of occupations ranked at the top to be more biased toward the male gender than the ones ranked later in the list. Similar to the above steps for obtaining the female-biased occupations, we update the embedding of human entities by taking the inverse of equation 5.1, and the occupations ranked at the top in the list are assumed as female-biased occupations. We have named the metric as embedding bias metric in later sections.

We note that other metrics like those proposed by Kediar et al. [99] have limitations. For instance, the metric Demographic Parity (DPD), as discussed by the authors, relies more on the bias included in the ground truth data than KG embedding. Further, this metric is heavily dependent on the classifier used and the number of classes (i.e., the number of occupations according to our setting) and, therefore, has limited power of explanation. In contrast, the metric we choose here is the most generalized one and reflects the biases that are introduced by the embedding learning algorithms.

5.1.2 Dataset

Wikidata is known as a free, open-source knowledge base that acts as knowledge storage for the structured data of sister Wikimedia projects such as Wikipedia, Wikivoyage, and others. Similar to other KGs, information is stored as facts or triples, containing a subject item, a property, and an object. Objects can be entities or literals such as a quantity, a string, etc. The subject/object items are denoted by URIs starting with 'Q' (e.g. Q5284 for *Bill Gates*), and properties are symbolized by URIs preceded by 'P' (e.g., P19 for *Place of birth*). We extract a specific set of triples from Wikidata based on a specific set of criteria and conducted experiments to find different trends of societal biases that exist in Wikidata.

Collection of geography dataset

For our work, we download the latest Wikidata dump[6] which is stored in a json format (bz2 format for compressed version). The dump consumes ~ 70 GB space in compressed form. We first converted it into standard KGTK[7] format for ease of data processing, which generated three files: a node file (~ 9.5 GB), an edge file (~ 189.5 GB) and a qualifiers file (~ 60.4 GB). KGTK is a standard Python library that facilitates easy manipulation of knowledge graphs. The node file contains the English labels, descriptions, and aliases of Qnodes and Pnodes. For example; Qnode Q943 has the label yellow, color as a description, and alias as the color yellow/Y/FFFF00. The edge file contains the triplets in the knowledge graph along with the details of the tail entity, like language, entity type, etc. Next, we filtered out the triplets that do not have either a wiki QID as the head entity or a wiki PID as the relation. Thus we end up with 1.32 billion triplets in the dataset with 93 million entities (QIDs) and 8763 relations (PIDs). Now, we aim to collect the triples corresponding to 13 geographies spanning over various continents which we have considered for our work - Arabia, Australia, Argentina, Brazil, France, Germany, India, Japan, Kenya, Russia, South Africa, United Kingdom, and the United States of America. The choice of the above geographies is motivated by the fact that the human entities

[6]https://dumps.wikimedia.org/wikidatawiki/entities/latest-all.json.bz2

[7]https://kgtk.readthedocs.io/en/latest/

covered by them are among the largest within their respective continents. This ensures large and diverse coverage of the dataset. The human entities belonging to a typical Geography are found by computing the overlap between the entities that are humans and those belonging to that specific Geography. To find this connection, we first extract the head entities of the collected triplets, which have the relation $P31$ (i.e., "instance of") and tail entity $Q5$ (i.e., "human"). Further, the head entities of triplets must be connected with the relation $P27$ (i.e., "country of citizenship", here indicating the specific Geography to which the entity belongs) and the tail entity as the corresponding QID of that country. Thus we gather the entities belonging to each of the geographies. In addition, we consider all the outgoing edges from these human entities and collect all the triplets with the head entity from the set of human entities we obtain. This collects the gender and occupation information of all the entities. In this way, Geography-specific knowledge graphs are created. For example, for constructing the Arabia dataset we consider outgoing edges from the human entities belonging to any country in the Arabian Peninsula. A brief statistic of entities, triplets, humans, and occupations is shown in Table 5.1.

Formation of a giant knowledge graph

To build a giant network joining the data of all 13 geographies, entities of individual Geography are combined together, which we denote by the set E. We construct the network by considering the edges among these entities. For this purpose, we extract all the triplets from the edge file with the head-entities and tail-entities that belong to our set of entities E and added to the graph. Finally, the giant subgraph of Wikidata that we construct has 22,254,967 triplets, 2,228,594 entities, and 894 relations. We use English labels obtained from the node file and qwikidata[8] to find the labels of the QIDs and PIDs. For our experiments, we extract the list of occupations that have at least one male and one female occurrence in a particular Geography under consideration.

[8]https://qwikidata.readthedocs.io/en/stable/

Geography	Triplets	Entities	Humans	occupations
Arabia	1,63,730	18,813	10,304	200
India	7,32,647	85,287	55,083	430
Japan	26,73,572	2,09,637	1,52,003	683
Russia	11,97,735	1,00,770	55,548	676
Australia	9,53,677	99,257	52,172	746
Kenya	45,224	6437	3220	100
South Africa	2,34,490	29,403	14,535	290
France	56,88,195	4,80,301	2,59,067	1272
Germany	56,81,167	4,09,367	2,42,894	1183
UK	38,26,106	3,46,240	1,56,601	1153
Argentina	6,14,872	59,858	34,770	420
Brazil	10,46,409	1,03,687	69,908	527
USA	1,00,53,843	7,34,686	4,26,281	1886

Table 5.1: Table showing statistics of entities, triples, humans and occupations in each of the 13 geographies.

5.1.3 Experiments

We investigate the impact of the sensitive attribute of gender on occupations in two possible ways, as mentioned earlier. First, we compute biases that exist in our KG dataset. Next, we measure biases on two graph embedding models across the geography dataset. For this experiment, we generate embeddings from scratch using our giant knowledge graph. Our experimental setup closely follows the bias measurement metric as that of [65] but is implemented on different geographies datasets that we sample from Wikidata. Precisely, followed by the embedding generation step based on the giant knowledge graph, we conduct bias measurement experiments separately on each demographic dataset.

Embedding methods	MRR	Hits@5	Hits@10	Hits@20
TRANSE	0.683	0.793	0.841	0.884
COMPLEX	0.774	0.939	0.978	0.992

Table 5.2: Link prediction result: evaluation of pretrained embeddings for TRANSE and COMPLEX

Generation of the knowledge graph embedding

As mentioned earlier, we employ two embedding learning models, namely TRANSE and COMPLEX in our experiment. These models are implemented using the similarity scores - L1 norm and dot product, respectively. In both cases, the dimension of the graph embedding is set to 100, considering the size of the network. We train the models by fixing the negative sample size (i.e., triples) to 3 and 10 for COMPLEX and TRANSE, respectively. We restrict the number of negative samples per positive triple because of the large size of our network. The training is completed using multiclass negative log-likelihood loss as the cost function and stochastic gradient descent as the optimizer. Finally, to test the quality of the trained embeddings, we evaluate them for the downstream task of link prediction using two standard metrics - Mean Reciprocal Rank (MRR) and Hits@n where n is set to 5, 10, and 20. In both metrics, a positive triple is ranked among 50 negative triples generated by permuting the subject/object side of the triples in the test set. For performing the evaluation, we randomly sample a test set of $10k$ triples from the knowledge graph and compute the two metrics for both TRANSE and COMPLEX. We perform three trials, each with a different set of randomly chosen $10k$ triples. The average results obtained for the three trials are noted in Table 5.2. We ensure that the triples in every sample test set are not seen during the training phase by the embedding algorithms. We have performed our experiments[9] with the help of the well-documented library AmpliGraph[10] on our dataset.

[9]https://docs.ampligraph.org/en/1.4.0/experiments.html
[10]https://docs.ampligraph.org/en/1.4.0/

Data bias in Wikidata

We study the data bias metric for all the geographies in our dataset. The metric assigns a bias score to every occupation that exists in the specific Geography. Further, on the basis of the bias score and threshold t, the occupations are categorized into three types: male-biased, female-biased, and neutral. Although the set of top-ranked male and female-biased occupations vary across geographies, we find the following list of occupations ranked at the top in almost all the geographies in our dataset as reported by the data bias metric.

- **male-biased occupations:** association football player, military personnel, politician, cricketer, etc.
- **female-biased occupations:** actor, writer, singer, model, etc.
- **neutral occupations:** epidemiologist, hydrologist, social scientist, intellectual, etc.

Bias measurement in KG embedding

We apply the embedding bias metric [65] (discussed briefly in the section 5.1.1) to our setting to measure the biases introduced by KG embedding methods. The metric assigns a bias score to every occupation that belongs to a specific Geography and generates two individual lists of *male* and *female* biased occupations for a particular Geography. Now, for every Geography, we rank the lists (i.e., male and female) of occupations according to their bias scores in descending order. To develop further insights, we analyze the ranked list of occupations (male and female) based on two orthogonal axes - (i) embedding learning methods and (ii) different geographies.

The goal of our experiments is to answer the following questions -

- How do the two embedding learning methods - TRANSE and COMPLEX rank

biased occupations for a given Geography? Are two rankings generated by two models similar or very different from one another?

- How do the rankings of gender-biased occupations vary across geographies? Do we observe any geographical differences among these ranked lists?

The similarity between different rankings is compared using the Jaccard similarity metric. This is done separately for each geography.

5.1.4 Results

Data bias vs algorithmic bias

Here we attempt to corroborate the first hypothesis - the biases present in the base data are altered by the embedding learning algorithms. For this purpose, we compute a rank list of male and female-biased occupations based on the (i) data bias metric and the (ii) bias metric for KG embedding. For the top K occupations based on the data bias metric, we attempt to find their corresponding ranks in the second-ranking list based on the embedding bias metric and compute the rank deviation. Let us assume that occupations p are ranked at r_1 and r_2 in the two rankings produced by the data bias and the embedding bias metrics, respectively. We take the inverse of the ranks and calculated rank deviation as follows - $(1/r_1) - (1/r_2)$ [11]. Finally, the overall rank deviation of the two rank lists is computed by averaging the individual rank deviation of the occupations that are ranked at top k positions based on the data bias metric. A positive (negative) value of rank deviation indicates that occupations are ranked at a lower (higher) position in the second rank list (i.e., bias metric for KG embedding) relative to the first one (i.e., data bias metric). The higher the deviation (on either positive or negative sides), the stronger the evidence that the inherent data bias is altered by the embedding learning algorithms.

[11] Simple $r_1 - r_2$ wouldn't differentiate between pairs $\{r_1 = 1, r_2 = 2\}$, and $\{r_1 = 100, r_2 = 101\}$ i.e reordering amongst the top-ranked vs low ranked occupations.

For a given Geography D, the rank deviation is computed for the following pairs of the ranked list of occupations ranked at K based on data bias metric.

- male-biased occupations by data bias metric *vs.* male-biased occupations by TRANSE's ranking using embedding bias metric.
- female-biased occupations by data bias metric *vs.* female-biased occupations by TRANSE's ranking using embedding bias metric.
- male-biased occupations by data bias metric *vs.* male-biased occupations by COMPLEX's ranking using embedding bias metric.
- female-biased occupations by data bias metric *vs.* female-biased occupations by COMPLEX's ranking using embedding bias metric.

The rank deviations for both male and female occupations across the geographies are noted in Table 5.3. The rank deviation values show that the two rankings generated by the data bias metric and the embedding bias metric are characteristically different for $K = 20$. Here the positive value of rank deviation indicates that the embedding learning methods ranked the top occupations according to the data bias metric at relatively lower positions. Further, almost all the geographies exhibited similar rank deviation. Thus we establish that the inherent data bias present in the dataset is indeed altered by the embedding learning algorithms. The results further indicate that if one discounts the inherent data bias, the embedding learning algorithms themselves introduce biases that can potentially affect the downstream NLP applications that rely on KG embeddings. Thus the pertinent question next is whether the biases introduced by the embedding learning algorithms are equivalent or different.

Comparison of two embedding methods: TRANSE with COMPLEX

In this section, we attempt to identify whether the biases introduced due to the embedding learning algorithms are the same or different across these algorithms. If they are indeed

Geography	TRANSE		COMPLEX	
	Male	Female	Male	Female
Arabia	0.15	0.13	0.11	0.12
India	0.17	0.13	0.17	0.10
Japan	0.16	0.11	0.16	0.10
Russia	0.16	0.12	0.12	0.15
Australia	0.16	0.15	0.11	0.14
Kenya	0.07	0.10	0.12	0.08
South Africa	0.10	0.14	0.14	0.16
France	0.15	0.13	0.17	0.10
Germany	0.18	0.09	0.11	0.17
UK	0.16	0.09	0.14	0.16
Argentina	0.13	0.13	0.12	0.12
Brazil	0.14	0.13	0.10	0.10
USA	0.18	0.16	0.16	0.14

Table 5.3: Rank deviation (as explained in section 5.1.4) of the two ranked lists generated by data bias metric and KG embedding bias metric truncated at the top (K = 20) ranked occupations.

different, then the downstream applications would differ based on the type of embedding algorithm used. For the two different categories of biased occupations (i.e., male-biased and female-biased), we compute the similarity between the two ranked lists of occupations generated by the embedding methods - TRANSE and COMPLEX. The comparison has been iterated over all the 13 geographies in our dataset and the result is shown in Table 5.4. Overall we find very little overlap between the two rankings of biased occupations obtained by the two embedding methods. Because of the difference in scoring functions, the embedding algorithms learn the underlying graph structure differently and perhaps influence the dissimilar ranking of our sample set of occupations. We compute this overlap for varying values of top K occupations in the rank list. For all values of K (20, 50, and 80), we observe a similar trend of minimal overlap with very low values of Jaccard similarity for both male and female-biased occupations. On a closer inspection into the list of bi-

ased occupations ranked by TRANSE and COMPLEX, we find that TRANSE pulls up very generic occupations at the top of the list. In contrast, Geography-specific occupations are ranked at higher positions in the list pulled up by the COMPLEX embedding. For example, a list of occupations ranked at the top 50 by the two independent embedding methods are -

- **male biased occupations ranked by TRANSE:** sportsman, specialist, climber, referee, leader etc.
- **male biased occupations ranked by COMPLEX:** adventurer, agent, charlatan, clergy, intelligence agent etc.
- **female biased occupations ranked by TRANSE:** activist, expert, discussion moderator, occupationsal, erudite etc.
- **female biased occupations ranked by COMPLEX:** feminist, traveler, faculty, home keeper, aviation employees etc.

While, in general, all overlaps are low between the biased rank lists generated by the two embeddings, we also observed that a few of the geographies, such as Germany, France, the UK, and the USA, have exceptionally low overlap (e.g., highlighted in red in Table 5.4). Certain other geographies like Arabia, Kenya, and South Africa have a relatively higher overlap between the biased rank lists generated by the two embeddings (highlighted in green in Table 5.4) in our dataset. The analysis reveals that the two embedding methods follow a different portfolio of biases in ranking biased occupations. In a nutshell, the sensitive attribute of gender has a varying impact on listing biased occupations when one employs different graph embedding models to embed the underlying knowledge graph.

Comparison of different geographies

Similar to the comparison of biased occupations based on embedding methods, we compare the rankings across the 13 geographies in our dataset. First, we generate the lists of

Geography	K=20		K=50		K=80	
	Male	Female	Male	Female	Male	Female
Arabia	0.11	0.14	0.16	0.22	0.34	0.31
India	0.03	0.08	0.08	0.12	0.09	0.16
Japan	0.05	0.03	0.03	0.06	0.05	0.13
Russia	0.03	0.03	0.01	0.08	0.03	0.14
Australia	0.03	0.0	0.03	0.03	0.05	0.11
Kenya	0.05	0.18	0.3	0.3	0.7	0.63
South Africa	0.05	0.05	0.1	0.18	0.11	0.21
France	0.00	0.03	0.02	0.04	0.02	0.08
Germany	0.00	0.00	0.01	0.05	0.01	0.07
UK	0.00	0.00	0.01	0.05	0.03	0.08
Argentina	0.05	0.08	0.04	0.14	0.07	0.17
Brazil	0.05	0.00	0.06	0.12	0.08	0.14
USA	0.00	0.00	0.01	0.03	0.01	0.05

Table 5.4: Jaccard similarity between the lists of biased occupations (male and female both) ranked by two different embedding techniques - **TRANSE** and **COMPLEX**. The results are obtained for the lists ranked at the top K (20, 50 and 80) biased occupations for all the geographies. The rows highlighted in red indicate very less similarity in the ranking of biased occupations obtained from the two embedding methods for all values of K. In contrast, the green highlighted rows for all values of K denote relatively slightly higher similarity of biased occupations ranked based on the two embedding methods.

male and female-biased occupations for each Geography using TRANSE and COMPLEX. This leads us to 4 different ranked lists for every Geography depending on the pair-wise combination of the sensitive attributes - male/female and two embedding methods. Formally, for a Geography D, the list of biased occupations L_1 (TRANSE-male), L_2 (TRANSE-female), L_3 (COMPLEX-male), L_4 (COMPLEX-female) are assumed to be ranked based on their bias scores and top 20 occupations are selected from each of the list for further analysis. For every Geography, we perform the same exercise leading us to 4 different lists for each of them. Now, for Geography D, we compared the 4 lists, namely L_1, L_2, L_3 and L_4 ranked at the top 20 with the corresponding lists of remaining

12 geographies. This produces a similarity index for each pair of the lists, thus generating 12 similarity values. Finally, we average these values and compute the mean and standard deviation for Geography D. The result is tabulated in Table 5.5.

We observe that average pairwise similarity across all the geographies for the ComplEx method is much lower than TransE. For example, the USA shows average an similarity of 0.32 and 0.15 with the remaining geographies for male-biased occupations when the embeddings are generated using TransE and ComplEx respectively. This phenomenon also holds for female-biased occupations. Further, we repeated the same procedure for the top occupations ranked at 50 and 80, and it produced a very similar trend. Thus, we believe it is a corollary of the fact that ComplEx pulls up very demographic-specific occupations (which are therefore different across geographies, resulting in low overlap) while TransE pulls up generic occupations common across the different geographies (resulting in a higher overlap).

Average similarity between geographies: The heatmaps in Figure 5.1 present an illustration of average similarity between every pair of geographies for the four combinations: TransE-male, TransE-female, ComplEx-male, ComplEx-female. The figures make it visually evident that TransE exhibits more similarity for a significant number of pairwise geographies compared to ComplEx. Further, the figures depict that female-biased occupations are more similar across geographies than male-biased occupations. This supports our insights presented in Table 5.5 (please refer to the columns named *female* for both TransE and ComplEx).

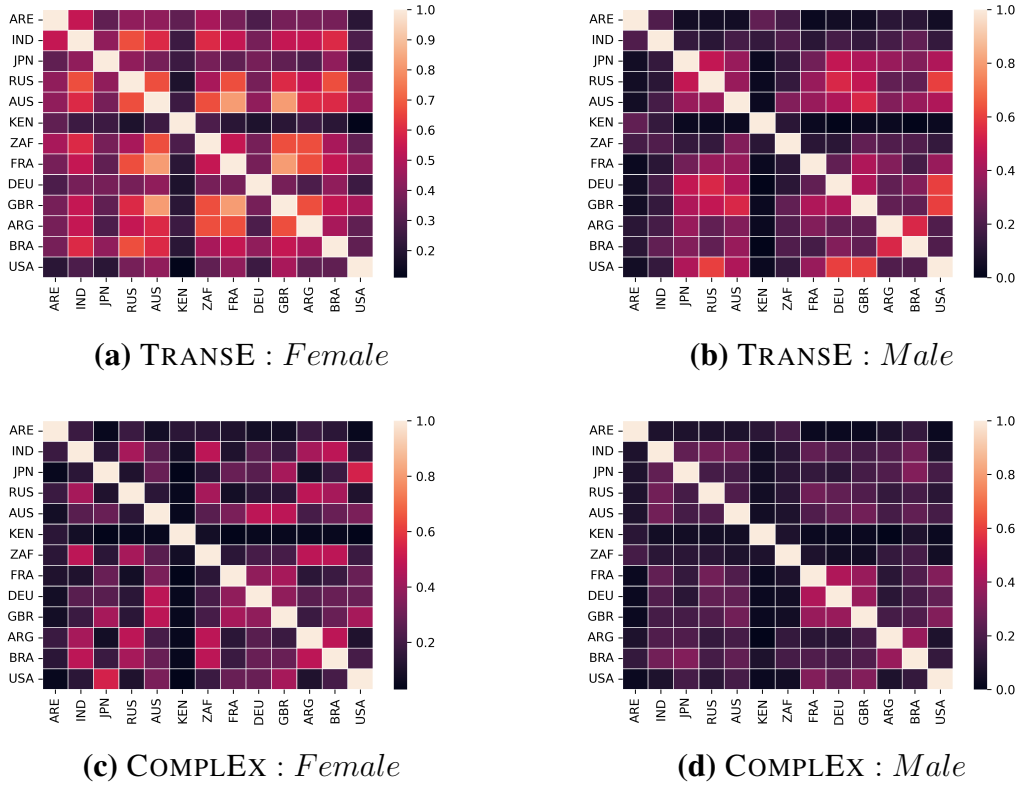
Most similar geographies: In addition, we carry out further analysis to find for a given Geography what are the topmost similar geographies (top three) across each of the combinations (i.e., TransE-male, TransE-female, ComplEx-male, ComplEx-female) presented in the heatmaps (please see Figure 5.1). The results are noted in Table 5.6. As has been observed so far, this list of the top three closest geographies varies widely based on the embedding method for each and every Geography. Further, for a particular embedding method and particular Geography, the top three closest geographies as per male-biased and female-biased occupations are also different. One interesting point

Geography	TRANSE				COMPLEX			
	Male		Female		Male		Female	
	mean	std. dev	mean	std. dev	mean	std. dev	mean	std. dev
Arabia	0.10	0.07	0.38	0.08	0.08	0.05	0.12	0.05
India	0.17	0.04	0.50	0.13	0.19	0.09	0.26	0.14
Japan	0.30	0.16	0.35	0.07	0.17	0.07	0.21	0.15
Russia	0.31	0.19	0.51	0.15	0.17	0.08	0.23	0.16
Australia	0.32	0.15	0.56	0.17	0.19	0.08	0.26	0.13
Kenya	0.06	0.07	0.23	0.05	0.05	0.03	0.06	0.03
South Africa	0.18	0.06	0.5	0.13	0.10	0.04	0.27	0.15
France	0.24	0.14	0.53	0.18	0.21	0.13	0.21	0.12
Germany	0.30	0.19	0.35	0.07	0.19	0.13	0.25	0.12
UK	0.32	0.18	0.54	0.18	0.20	0.12	0.27	0.15
Argentina	0.26	0.13	0.48	0.15	0.15	0.09	0.26	0.16
Brazil	0.25	0.13	0.47	0.12	0.21	0.08	0.29	0.14
USA	0.32	0.21	0.32	0.10	0.15	0.10	0.23	0.15

Table 5.5: Average similarity (mean, std. deviation) across geographies for biased occupations (male and female individually) ranked at top 20. The similarity has been computed for both the embedding methods TRANSE and COMPLEX.

that we note is that the UK features in the list of top three for all four combinations in the case of at least three different geographies – USA, Australia, and France. Manual inspection shows that the occupations that connect the UK to these geographies are rugby union player (male), rabbi (male), scientific illustrator (female), and suffragist (female), which possibly corresponds to the shared cultural heritage and economic parity of these geographies with the UK. In Table 5.7, we note the most similar pair of geographies and manually inspect whether this similarity can be attributed to either socio-cultural or geographical closeness. TRANSE pulls up pairs like (UK, USA), (Australia, UK), (India, South Africa) as similar. These pairs are geographically quite apart, sometimes even appearing in opposite hemispheres. The reason for their similarity can be primarily

attributed to their close socio-economic and socio-cultural complexion. Since TRANSE pulls up biased occupations that are generic, this socio-economic and socio-cultural closeness gets naturally manifested. All the Geography pairs that we manually find to be close in this dimension are marked in red in Table 5.7. On the other hand, certain pairs that are geographically close are pulled up as similar by COMPLEX. Closer geographies should have similar types of nuances in their list of most biased occupations, which is what is extracted by COMPLEX and hence the similarity (observed more across the male axis compared to the female axis). We mark all these cases in blue in Table 5.7.



(a) TRANSE : $Female$ **(b)** TRANSE : $Male$

(c) COMPLEX : $Female$ **(d)** COMPLEX : $Male$

Figure 5.1: Heatmaps showing similarity between different geographies for the male and female-biased occupations ranked at $K = 20$. The heatmaps in the top and bottom rows are generated for TRANSE and COMPLEX respectively. The abbreviations used for the names of the countries are as per the ISO 3166 standard.

5.1.5 Discussion

The primary objective of this work is to bring forth the fine-grained details of biases manifesting from the disparity of data representation across the different geographies as well as the non-neutral learned embeddings from knowledge graphs. In this section, we discuss a few other interesting insights that we obtained from our analysis.

Unique occupations per Geography

In an ideal scenario, occupations should be free from socio-cultural and geographic delineations. We revisit the top listed occupations (ranked at 100) for each Geography and tried to find unique occupations that signify Geography-specific characteristics. A closer inspection of this list shows that the occupations pulled up by ComplEx embedding resemble nuanced examples of Geography-specific characteristics. For example, the occupations *sitarist* for India, *Islamic jurist* for Arabia, *bhikkhu* for Japan symbolize the cultural heritage of a specific geographic area. In contrast, the occupations selected by TransE mostly represent different sports-related occupations for males and media-related occupations for females. Moreover, some controversial occupations such as *playboy playmate* (USA), *pornographic film director* (Russia), *witch* (Kenya), and *cunning folk* (UK) emerge as female-dominated occupations and *terrorist* (Arabia), *militant* (Arabia), *holocaust denier* (Germany) and *anarchist* (Argentina) emerge as male-dominated occupations. These need to be judiciously audited by the Wikipedia community at large in order to promote a safer cyberspace.

Female biased occupations are less diverse

Although the Wikimedia Foundation provides an open-source gender-neutral editing environment to all the digitally literate people around the globe, it is now well-known that there exists a substantial gender gap in many Wikimedia projects in terms of the number of editors and editing ractices [45, 108, 187]. Further, researchers [180] discovered

Geography	TransE		ComplEx	
	Male	Female	Male	Female
Arabia	Kenya, India, South Africa	India, South Africa, Russia	South Africa, Brazil, Kenya	India, Russia, Argentina
India	Brazil, Arabia, South Africa	Russia, South Africa, Brazil	Russia, Brazil, Japan	South Africa, Brazil, Russia
Japan	Russia, Germany, UK	India, Russia, Brazil	Brazil, India, Argentina	USA, UK, Australia
Russia	USA, Germany, Japan	India, Australia, France	India, France, Germany	Argentina, India, South Africa
Australia	UK, Germany, Japan	France, UK, Russia	UK, India, Germany	Germany, UK, France
Kenya	Arabia, India, South Africa	Arabia, South Africa, India	Arabia, South Africa, Brazil	Arabia, India, South Africa
South Africa	Australia, Brazil, Argentina	Australia, UK, Argentina	Arabia, Brazil, Argentina	India, Argentina, Brazil
France	UK, USA, Russia	Australia, UK, Russia	Germany, UK, USA	UK, Germany, Australia
Germany	USA, Russia, Japan	Australia, India, Japan	France, UK, USA	Australia, France, UK
UK	USA, Australia, Russia	Australia, France, South Africa	Germany, USA, Australia	Australia, Japan, USA
Argentina	Brazil, Japan, Australia	South Africa, France, UK	Brazil, India, Japan	Russia, South Africa, Brazil
Brazil	Argentina, Australia, Japan	Russia, India, Australia	Argentina, Japan, India	India, South Africa, Argentina
USA	UK, Germany, Russia	UK, Australia, France	UK, France, Germany	Japan, UK, Australia

Table 5.6: Table showing geographies that exhibit similarity in ranking biased occupations (ordered in decreasing value of similarity in each cell).

Geography	Male	Female
TRANSE	(UK, USA), (Australia,UK), (Russia, Japan), (Russia, UK)	(France, UK), (India, Russia), (India, South Africa), (Australia, UK)
COMPLEX	(France, Germany), (France, UK), (Germany, UK), (Argentina, Brazil)	(Australia, Germany), (Australia, UK), (India, South Africa), (Argentina, Brazil)

Table 5.7: Table showing most similar pairs of geographies for the combinations- TRANSE-male, TRANSE-female, COMPLEX-male, COMPLEX-female.

implicit gender biases in describing human biographies. Similar to other Wikimedia projects, gender inequality and underrepresentation of content about females compared to males exist in Wikidata as well in varying degrees. In fact, a recent study [205] finds that only 22% of Wikidata items that represent people are about women. Likewise, our detailed analysis has identified several instances in which male-biased occupations exhibit a diverse range of occupations (see Figure 5.2b) across different geographies. In contrast, female-biased occupations are confined to similar types of occupations, mostly, for example, occupational, activist, erudite, etc. (see Figure 5.2a). Further, the biased occupations in each type (male and female) are represented in such a way that it reflects the so-called gender discrimination that persists in our society. For example, females are categorized into a typical generic set of occupations, such as activist, specialist, home keeper, etc., whereas males are subjected to the role specific occupations, e.g., leader, vehicle operator, believer, etc. (as depicted in Figure 5.2). To quantify this gender disparity with regard to the representation of occupations across geographies, we measured the diversity of individual groups of male and female-biased occupations as follows.

First, we collect the list of the top 50 biased occupations from each of the geographies

and combined them into a set of the distinct vocabulary of occupations separately for the male and the female categories. We calculate the probability (p_{prof}) of the occurrence of each of the biased occupations ($prof$) in the 13 geographies of our dataset. Now, we calculate the Shannon entropy for each of the occupations ($-p_{prof} \log p_{prof}$) and finally sum over all the occupations ($-\sum_{prof} p_{prof} \log p_{prof}$). This process is iterated on both male and female categories of biased occupations for the two embedding methods separately. A higher entropy value would indicate a larger diversity of occupations over the geographies. The obtained entropy values are shown in Table 5.8. The entropy measure clearly illustrates that the male-biased occupations exhibit a larger variety than female-biased ones for both the embedding generation methods. Perhaps Wikidata editors tend to add certain types of occupations that substantially favor masculinity over feminism, which in turn reflects the real-world gender bias.

Gender	TRANSE	COMPLEX
Male	55.69	68.98
Female	41.96	50.32

Table 5.8: Table showing entropy-based diversity for representing distinct male and female-biased occupations.

(a) Word cloud: Female

(b) Word cloud: Male

Figure 5.2: Word clouds showing distinct (a) female-biased and (b) male-biased occupations.

5.1.6 Summary

In this work, we perform a large-scale audit to investigate the presence of data bias in the knowledge graph (i.e, Wikidata) by leveraging a dataset comprising 13 geographies. Further, we also study the presence of algorithm bias in KG embedding by applying two distinct embedding learning models – TRANSE and COMPLEX. We show that diversity in the dataset and embedding methods with regard to specific sensitive features (e.g., the impact of gender on occupation) can result in stark differences in the ranking of biased occupations.

5.2 Bias and fairness in link prediction task in Wikidata

In spite of rampant utilization of Knowledge graphs in various applications, e.g., language model development [3, 128], question answering [86, 170], personalized recommendation [192] and others, it is widely acknowledged that even the largest and most comprehensive KGs have limitations in terms of incompleteness. They only contain a fraction of the vast amount of real-world knowledge that they ideally should encompass. To address this issue of incompleteness, link prediction (LP) techniques utilize the existing facts in KG to make predictions about new facts. LP models commonly make use of vectorized representations of entities and relations, known as embeddings, which are learned through some learning models and further utilized in predicting the missing link between entities. These models have achieved state-of-the-art performances in the task of link prediction [165, 188].

Bias versus fairness: Unfortunately, while high-quality structured content is a plus, a wide range of societal and human biases are inherent to KGs in many ways – be in the form of sampling strategy or the judgmental view. A pertinent question in the research community exists regarding the source of biases in knowledge graphs [55]. It is important to recognize the interplay between the biases that are inherently present in KGs and those that emerge during the embedding learning process. In the previous work, we see

that the data bias existing in the knowledge graph from its inception can undergo further alterations during the embedding learning phases, specifically due to the characteristics of the learning algorithms, referred to as algorithm bias. This can lead to biased predictions in many downstream applications. In this context, it is worth mentioning that KGs indeed face challenges when it comes to balancing bias versus fairness. To address this issue, researchers have proposed various methods for auditing and improving KGs [111]. In addition to detecting and addressing biases, researchers have also looked into the impact of biases on tasks such as link prediction used for KG completion. Through a systematic analysis, authors in [165] identified three types of data biases that can affect link prediction datasets and models. Another research has investigated the issue of structural imbalances in KGs, particularly in the context of KG completion [177].

Our contribution: In this work, we consider a large corpus extracted from a knowledge graph (Wikidata), comprising knowledge triples from 21 geographies spanning multiple continents and investigate the biases that creep in while learning representations using various algorithms. Our study focuses on link prediction as a downstream task to measure the biases in Wikidata. We pose two key research questions to align our contributions in seeking answers–

- **RQ1:** How do social biases (i.e., gender: male/female and age: young/old in our work) impact predictive outcomes related to an observable (occupation or profession in our case) in the link prediction task given a knowledge graph?
- **RQ2:** How do biases vary by varying the geo-social data used in downstream link prediction, and whether the patterns of biases so revealed point to some universal characterization of the geographies?

We extract triples from 21 different geographic locations around the globe, including Arabia, India, Israel, Japan, South Korea, Turkey, Russia, Australia, New Zealand, Egypt, Nigeria, South Africa, France, Germany, Spain, United Kingdom, Argentina, Brazil, Canada, Mexico and the United States of America. To investigate how social biases affect the fair link prediction of occupations, we propose a noble framework named **AuditLP**

to measure the biased predictions given the sensitive attributes of human entities. Our framework is built upon popular knowledge graph embedding learning (KGE) algorithms – TRANSE [30], DISTMULT [201], COMPGCN [173], and GEKC [127] which generate knowledge graph embeddings as the features to be used in the prediction. The framework is designed to hide relationship links (between human entities and occupation names) from the training instances while learning the embeddings by the KGE algorithms. Next, the learned embeddings, aka features, are passed to a classifier to predict the links. Finally, the predictions are analyzed in light of the notion of fairness. Analyzing the prediction outcomes by *AuditLP*, we came across the following key findings –

- The classifier outcomes are unfair regarding sensitive attributes for a given set of professions.
- Our study also reveals that the choice of data, i.e., geographies, significantly impacts the variance of biases. Specifically, the social biases present in different geographies manifest into a clear partition of the world into two distinct regions: the Global North and the Global South, characterized by their different geo-social and economic attributes[12]. Surprisingly, this result is true for all the algorithms we used in *AuditLP* – TRANSE, DistMuslt, COMPGCN, and GEKC– even though their inner workings are quite different. Such a result indicates that this observation has a universal underpinning.

5.2.1 Dataset

We extract a specific set of triples from Wikidata based on certain criteria and finally obtain the carefully curated dataset.

Curation of geography dataset: In our research, we make use of the latest Wikidata dump[13], which is stored in JSON format, compressed in the bz2 format, taking up

[12]https://en.wikipedia.org/wiki/Global_North_and_Global_South

[13]https://dumps.wikimedia.org/wikidatawiki/entities/latest-all.json.bz2

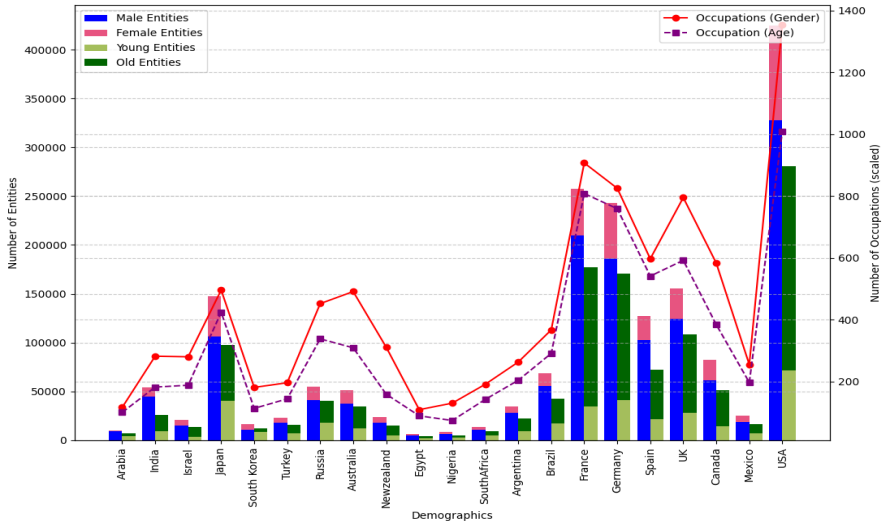


Figure 5.3: Figure showing the count of human entities (i.e., male/female, young/old, and corresponding number of occupations per geography in our dataset.

approximately 70 GB when compressed. Initially, we process this dump through the widely-used KGTK[14] library to streamline data handling. This step results in the creation of three distinct files: a node file (~ 9.5 GB), an edge file (~ 189.5 GB), and a qualifiers file (~ 60.4 GB). KGTK is a Python library commonly employed for simplifying the manipulation of knowledge graphs. Within the node file, one can find information like English labels, descriptions, and aliases linked to both Qnodes and Pnodes. In contrast, the edge file contains the triplets found within the knowledge graph, along with additional details about the tail entity, such as its language and entity type, etc. Subsequently, we undertake a filtration process to exclude triplets that lack a Wikidata identifier (QID) as the head entity or a Wikidata property identifier (PID) as the relation. This filtration results in a dataset comprising an impressive 1.32 billion triplets, encompassing 93 million entities (QIDs) and 8,763 relations (PIDs). Our primary aim is to collect the triples associated with 21 geographic attributes spanning various continents, which we have specifically chosen for our research. These geographies include Arabia, Australia, Argentina, Brazil, Canada, Egypt, France, Germany, India, Israel, Japan, Mexico, New Zealand, Nigeria, Russia, Spain, South Africa, South Korea, Turkey, the United Kingdom,

[14]https://kgtk.readthedocs.io/en/latest/

and the United States of America. To identify human entities within a specific geography, we compute the overlap between entities that are both human and associated with that particular geography. To establish this connection, we extract the head entities from the collected triplets that have the relation $P31$ ("instance of") and the tail entity $Q5$ ("human"). Moreover, the head entities of these triplets must be linked to the corresponding country using the relation $P27$ ("country of citizenship") and the tail entity representing the QID of that specific country. By doing so, we gather the human entities belonging to each geography. In addition, we consider all outgoing edges from these human entities and collect all the triplets where the head entity belongs to the set of human entities we obtained earlier. This enables us to collect gender and occupation information for all the entities. For every human entity, we extracted their date of birth from the Wikidata triples and computed their age as per the date of the experiment (i.e., December 2023). Next, we divided the human entities in every geography into two age groups – 'young' and 'old'. The persons whose ages are in the range of 19–45 years are identified as young. On the other hand, people, having age in the range 60–90 years are labeled as old. We kept a gap of 15 years between the two ranges so that the distinction between the young and old is very prominent. This process allows us to create geography-specific knowledge graphs. For the Arabia dataset, we specifically consider the outgoing edges from human entities belonging to any country within the Arabian Peninsula. A brief statistics of humans – male and female (young and old) and the corresponding number of professions are shown in Figure 5.3. The figure clearly illustrates that there is an imbalance in the proportion of male and female entities in each geography. This difference is the result of sampling bias that mimics the inequalities found in the actual world. Further, the distribution of professions in America or Europe is in stark contrast to those from Asia or Africa. This difference highlights how Western culture dominates other cultures on collaborative platforms like Wikidata. It is important to note that we were careful while experimenting with our proposed framework to maintain the diversity present in the geographic dataset of Wikidata. Our dataset is carefully designed to accomplish the downstream task of link prediction and measure biases in the task. The uniqueness of the large dataset lies in the two factors– **(1)** categorization of knowledge triples based on geographies and **(2)** availability of sensitive information, such as age and gender, about human entities.

5.2.2 Background

Definition of bias: Understanding of bias in link prediction is grounded in the concept of representational harm, specifically referring to the differences in system performance for different social groups. For instance, in the case of gender bias in predicting occupations, if a link prediction model is less accurate in predicting the occupations of women compared to men, it would be considered to exhibit harmful behavior. This is because deploying such a biased model in downstream applications could reduce its utility for women compared to men, potentially reinforcing societal inequalities in the predictive performance of present-day AI systems, including the large language models.

Notion of fairness: The notion of fairness refers to the concept of treating individuals and groups equitably in different aspects, including algorithms, predictive outcomes, etc., in general by the whole automation system. In the context of link prediction, fairness denotes that the predictions should not systematically discriminate against particular individuals or groups based on certain sensitive attributes such as gender, race, religion, age, etc. We replicate this notion of fairness in our experiment by assessing the impact of sensitive attributes of a human entity, such as gender or age, on the accuracy of predictions for target properties, like occupation. Let us assume, in a classification setup for predicting whether a human has a specific occupation or not, the following notations stand for–

1. G: Protected or sensitive attribute for which (non)discrimination should be established. In our experiment, it is – (a) gender: male and female (binary division), (b) age: young and old.
2. Y: The actual class (1 or 0 in our case) labels as per the dataset, i.e., whether a particular human entity has the occupation p (label = 1) or not (label = 0).
3. $\overline{Y}$: Predicted outcome, i.e., class labels for an individual obtained from the classifier.

Now, the predictive outcomes of the classifier considered for identifying fair classification

are mentioned below. Here, m and f stand for male and female, respectively. On the other hand, young and old people are denoted by yn and ol, respectively.

True positive rate (in case of gender):

$$TPR_m = P(\overline{Y} = 1|Y = 1, G : gender)$$

$$TPR_f = P(\overline{Y} = 1|Y = 1, G : gender)$$

True positive rate (in case of age):

$$TPR_{yn} = P(\overline{Y} = 1|Y = 1, G : age)$$

$$TPR_{ol} = P(\overline{Y} = 1|Y = 1, G : age)$$

False positive rate (in case of gender):

$$FPR_m = P(\overline{Y} = 1|Y = 0, G : gender)$$

$$FPR_f = P(\overline{Y} = 1|Y = 0, G : gender)$$

False positive rate (in case of age):

$$FPR_{yn} = P(\overline{Y} = 1|Y = 0, G : age)$$

$$FPR_{ol} = P(\overline{Y} = 1|Y = 0, G : age)$$

We followed two state-of-the-art fairness metrics– *Equal Opportunity* and *Equalised Odds* that measure the fair prediction of a classification task, i.e., link prediction in our experiment.

- **Equal opportunity**: It demands the positive outcome to be independent of the protected attribute, G, conditional on Y being an actual positive. In other words, the True Positive Rate (TPR) is to be the same for each element of the protected attribute, e.g., male and female or young and old in our classification setup.

$$TPR_m = TPR_f$$

$$TPR_{yn} = TPR_{ol}$$

- **Equalized odds**: It defines the positive outcome to be independent of the protected attribute G, conditional on the actual Y. In terms of classification outcomes, the True Positive Rate (TPR) and False Positive Rate (FPR) are to be the same for each element of the protected attribute, e.g., male and female or young and old in our classification setting.

$$TPR_m = TPR_f \; and \; FPR_m = FPR_f$$

$$TPR_y = TPR_o \; and \; FPR_{yn} = FPR_{ol}$$

Based on the above definitions, we label an occupation as either male-biased/female-biased/gender-neutral or young-biased/old-biased/age-neutral, which is introduced in the subsequent sections.

KGE models: Using our proposed framework, AuditLP, we demonstrate social biases in link prediction. To achieve this, we select four widely used embedding learning algorithms – TRANSE, DISTMULT, COMPGCN, and GEKC to generate knowledge graph embeddings for the triples in our dataset. The rationale behind choosing these models is that each of them represents a popular genre of embedding learning, and our objective is to investigate the performance of the algorithms of each genre. Our hypothesis is that the other algorithms from each of these genres would behave similarly to the representative ones we chose. According to the standard categorization of KGE models [191], TRANSE [30] is a translational approach, DISTMULT [201] distance-based semantic matching approach, and COMPGCN [185] is convolution neural network based approach. GEKC [127] is the latest addition to the KGE learning paradigm, which uses generative KGE circuits to enhance the efficiency and reliability of triple predictions for the missing link prediction task. Our study aims to demonstrate how models from each of these genres perform on our curated dataset.

5.2.3 AUDITLP: bias measurement framework

To investigate how social biases are encoded in KGs and are further perpetuated to downstream tasks, we proposed our framework named AUDITLP for predicting the links in a given geographic dataset. The framework comprises four main stages, and the steps are further detailed in the following sections. The methodology is described in terms of the sensitive attribute of gender and its division of gender-biased occupations– male-biased, female-biased, and gender-neutral. An exactly similar approach is followed when measuring biased outcomes in terms of sensitive attribute age, in which the occupations are grouped into three categories– young-biased, old-biased, and age-neutral.

Hiding edges: Each geography present in our dataset has occupations with at least

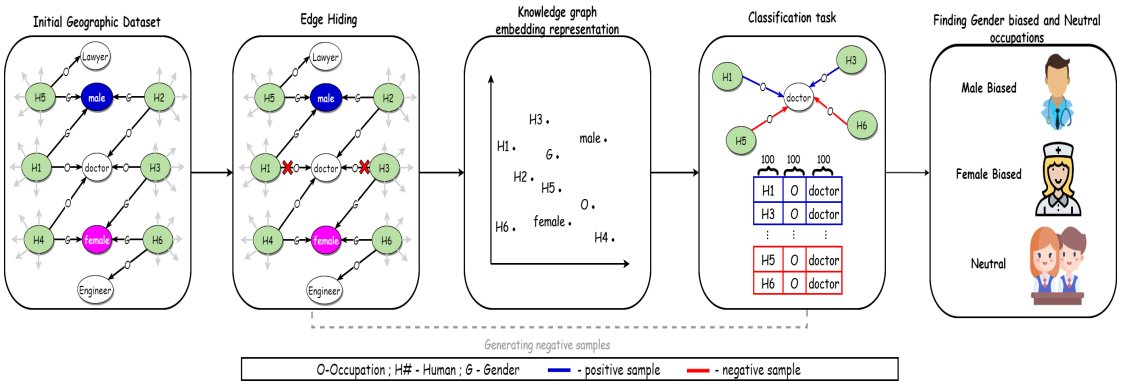


Figure 5.4: Schematic for our experiments showing different steps in case of sensitive attribute gender– edge hiding, embedding generation, and classification of occupations.

two male and two female human entities, which are considered in this step. Since our goal is to predict the occupation of a human entity, the occupation information should not be encoded in the human entity embedding while generating it in the training phase. Therefore, in this step, we hide (i.e., remove) the occupation information of the human entities whose occupations have to be predicted in the classification phase. Let us assume that there exists x number of human entities having occupation p in a geography d, with male and female entities as m and f, respectively. The triple $< H_1$, *occupation*, $O_1 >$ indicates that the human entity H_1 has the occupation O_1. We randomly select 50% of the human entities from x while maintaining the male-female ratio and remove their

occupation (triples similar to $< H_1$, *occupation,* $O_1 >$) to get the filtered dataset. The occupation triples removed from the initial graph are treated as positive samples for classification. Further, an equal number of negative samples are generated from the human entities in the original graph that do not have the occupation p.

Generation of knowledge graph embeddings: We train the filtered geographic by the embedding learning models– TRANSE, DISTMULT, COMPGCN, and GEKC one at a time. For the learning algorithms (except GEKC), we set the dimensions of the graph embedding to 100 and 200 for GEKC. The models are trained for a varying number of epochs depending on the size of the dataset using a fixed negative sample size of 10, with multiclass negative log-likelihood loss as the cost function and SGD as the optimizer. We utilize the python library AmpliGraph[15] for implementing and generating embedding by TRANSE and Pykeen[16] in case of DISTMULT and COMPGCN. For GEKC, we utilized the public repository[17] made by the authors to train the geographies. We evaluated the embeddings in terms of standard metrics like MRR and Hits@n (n is set to 5, 10, and 20) and achieved results that are comparable to benchmark results reported in the literature [7].

Predicting occupation of human entities: After generating embeddings of humans, relations, such as 'has_occupation', and occupation entities, our goal is to predict the occupations of human entities. Given a triple of the form $< H_1$, *has_occupation,* $O_1 >$, the classifier answers whether the human entity H_1 has the occupation O_1 or not. In other words, the link prediction classifier takes as input a concatenation of vectors representing – (gendered) head entity vector $\odot$ relation entity (i.e., 'has_occupation') vector $\odot$ tail entity (i.e., occupation name) vector and predicts a 0 (no relation exists) or 1 (relation exists), based on a set of training instances. For training purposes, positive triples are the triples that have occupation p in the initial dataset, while negative triples are selected from the human entities that do not have occupation p. The dimension of the triples is kept at 300, which is obtained by concatenating the embeddings of the human, relation, and occupation entities, each having 100 dimensions. To accomplish the classification,

[15]https://docs.ampligraph.org/en/1.4.0/

[16]https://pykeen.readthedocs.io/en/stable/

[17]https://github.com/april-tools/gekcs

we use an MLP classifier, and an 80:20 split was used between the training and test sets. The classification result, i.e., the true positive rate and the false positive rate, are considered for further analysis. For each embedding algorithm, we perform classification for each of the geographies individually. To measure the classification results based on gender divisions – male and female, we calculate TPR and FPR for both male and female entities individually, as defined in the previous section.

Finding gender-biased occupations: Now, we attempt to categorize the occupations as gender-dominated or gender-neutral based on the TPR and FPR values as predicted by the classifier. Utilizing the two metrics of fairness– *Equal Opportunity* and *Equalized Odds* (as discussed in the section 5.2.2), the occupations in every geography are identified to be in one of three categories – male-biased, female-biased, and gender-neutral.

An occupation o is identified as male-biased if it satisfies any one of the two definitions –

$$TPR_m(o) > TPR_f(o) \tag{5.3}$$

$$TPR_m(o) = TPR_f(o), FPR_m(o) > FPR_f(o) \tag{5.4}$$

Similarly, a female-biased occupation follows any of the two definitions below –

$$TPR_f(o) > TPR_m(o) \tag{5.5}$$

$$TPR_m(o) = TPR_f(o), FPR_f(o) > FPR_m(o) \tag{5.6}$$

Naturally, an occupation o is said to be gender-neutral if it satisfies the following constraint.

$$TPR_m(o) = TPR_f(o), FPR_f(o) = FPR_m(o) \tag{5.7}$$

We observe that sometimes the TPR and FPR values are nearly equal for male and female entities for certain occupations in the test dataset, and those are ignored from being categorized as male/female-biased occupations. Henceforth, we consider an occupation as *male-biased* if it satisfies any of the following.

$$TPR_m(o) - TPR_f(o) \geq t_1 \tag{5.8}$$

$$TPR_m(o) = TPR_f(o), FPR_m(o) - FPR_f(o) \geq t_2 \tag{5.9}$$

and *female-biased* if any of the following holds true.

$$TPR_f(o) - TPR_m(o) \geq t_1 \tag{5.10}$$

$$TPR_m(o) = TPR_f(o), FPR_f(o) - FPR_m(o) \geq t_2 \tag{5.11}$$

For *gender-neutral* occupations, we considered the following.

$$|TPR_m(o) - TPR_f(o)| < 0.01 \quad |FPR_m(o) - FPR_f(o)| < 0.01 \tag{5.12}$$

To set the threshold t_1, we take the differences of TPR_m and TPR_f for all the occupations and compute $\mu - \sigma$ of these differences. Similarly, we compute t_2 based on the FPR differences. Such methods of eliminating outlier values have been abundant in the literature [164]. If both t_1 and t_2 are as low as 0.01, we denote such cases as gender-neutral occupations for all geographies alike. Thus, following this pipeline, we have three lists of occupations prepared for every geography – (i) male-biased, (ii) female-biased, and (iii) gender-neutral. In the case of sensitive attribute age, we set $\mu + \sigma$ of the TPR and FPR differences as the corresponding thresholds t_1 and t_2 to cut off all the outlier cases. An occupation is assumed as age-neutral if the absolute differences of TPR_{yn} and TPR_{ol} or FPR_{yn} and FPR_{ol} is lesser than 0.01. In the case of sensitive attribute age, we divide the lists of occupations into three categories of biased occupations– (i) young-biased, (ii) old-biased, and (iii) age-neutral.

Metric	TRANSE		DISTMULT		COMPGCN		GEKC	
	μ	σ	μ	σ	μ	σ	μ	σ
TPR_m	0.90	0.09	0.84	0.09	0.85	0.04	0.90	0.09
TPR_f	0.86	0.10	0.83	0.08	0.85	0.05	0.81	0.09
FPR_m	0.15	0.11	0.17	0.14	0.13	0.03	0.07	0.03
FPR_f	0.16	0.10	0.20	0.13	0.13	0.04	0.07	0.04
Acc_{gender}	86.50	8.43	82.66	10.99	81.59	18.70	92.97	1.24
$F1_{gender}$	0.86	0.8	0.80	0.12	0.85	0.03	0.89	0.04
AUC_{gender}	0.93	0.8	0.86	0.14	0.94	0.02	0.92	0.05
TPR_y	0.87	0.11	0.85	0.14	0.82	0.05	0.92	0.04
TPR_o	0.85	0.10	0.83	0.16	0.83	0.04	0.88	0.03
FPR_y	0.15	0.14	0.18	0.16	0.16	0.04	0.06	0.03
FPR_o	0.19	0.13	0.22	0.18	0.14	0.04	0.04	0.01
Acc_{age}	84.32	0.10	92.18	0.12	84.03	3.79	93.01	0.03
$F1_{age}$	0.84	0.09	0.92	0.13	0.84	0.04	0.90	0.03
AUC_{age}	0.91	0.10	0.97	0.13	0.92	0.03	0.94	0.04

Table 5.9: Table showing the mean and standard deviation of different metrics averaged over all geographies for TRANSE, DISTMULT, COMPGCN, and GEKC. Here, the first **seven** rows tabulate metrics computed for gender, and the last **seven** rows are for age.

5.2.4 Results

Fairness in classification

We run the classifier on each of the 21 geographies individually and iterated the process for TRANSE, DISTMULT, COMPGCN, and GEKC. We obtain accuracy (aka Acc), F1-score, AUC score, TPR_m, TPR_f, FPR_m, FPR_f for each occupation, and average the results over all of the occupations and all geographies. We report this average (μ) along with the standard deviation (σ) for each metric in the case of the training accomplished by TRANSE, DISTMULT, COMPGCN, and GEKC in Table 5.9 (first seven rows). Similarly, for age, we compute accuracy (aka Acc), F1-score, AUC score, TPR_{yn}, TPR_{yn},

FPR_{ol}, FPR_{ol} for each occupation and the average value over all the occupations and geographies and corresponding standard deviation are tabulated in the last seven rows of Table 5.9. Overall results show that at an aggregate level, i.e., considering all the geographies together, the average values of TPR and FPR feature slight differences when the males and females are compared. This is true for the age attribute as well. Therefore, it is clear that the prediction outcome is not fully fair and equal across the social groups. However, the picture changes when we observe the occupations at the individual geography level. We utilize equations (5.8) through (5.12) to create a list of occupations categorized as male-biased, female-biased, or gender-neutral for each of the 21 geographies. The biased occupations obtained in the case of each attribute are placed in the Tables below.

- Table 5.13 (A) and (B) list the professions marked as gender-biased (male/female) and neutral for the algorithms TRANSE and DISTMULT.
- In the Tables 5.14 (A) and (B), the professions denoted as gender-biased or age-biased in case of embedding generated by COMPGCN are tabulated.
- The professions labeled as age-biased (young/old) by the embedding algorithms– TRANSE and DISTMULT are placed in the Tables 5.15 (A) and (B).
- Professions that are labeled as biased for the attributes gender, and age in case of embedding generated by GEKC are placed in the Tables 5.16 (A) and (B).

We have come across various interesting observations where occupations are biased differently in different geographies. For instance, in the USA, occupations like real estate brokerage are male-biased, while occupations like prostitution and pornographic actors are female-biased. We observe that countries like France, Germany, and Spain have the occupations of linguist, engineer, doctor, professor, and economist as male-biased, and the female-biased professions include sports like volleyball, alpine skier, blogger, etc. In contrast, we note that likewise USA, in Germany too, pornographic actor is identified as a female-biased occupation. Further, a few occupations, such as film actors and artists, manifest as gender-neutral in many countries, such as Argentina, Brazil, South Africa, Canada, Mexico, etc. A similar trend is observed in the case of the sensitive attribute age.

Laborious occupations, such as tennis player, boxer, war photographer, police officer, etc., are categorized as young-biased across almost all the geographies – Arabia, South Africa, Nigeria, Canada, Spain, UK, USA, etc. On the other hand, sports commentators, computer scientists, film actors, vocalists, and other less labor-intensive occupations are labeled as old-biased occupations. Further, some age-neutral occupations in different countries include basketball players in the USA, researchers in France, songwriters in India, dub actors in Israel, poets in Mexico, etc. Although geographies with a larger number of triples, such as the USA, Germany, and France, contain occupations in almost all categories in contrast to smaller geographies such as Arabia, Israel, etc., overall, there is a wide variance in the biased occupations for both gender and age across all these geographies. Occupations categorized (i.e., male/female, young/old) by the fairness metrics give a very detailed picture of the biased professions across geographies, and it is not easy to see a universal pattern of social biases. Therefore, we next outline an approach to automatically extract such patterns.

Social biases at macro level

To discover universal patterns, we characterize each geography based on a five-dimensional vector representation. Each entry in this vector corresponds respectively to each of the conditions described in equations (5.8)-(5.12). For a given occupation if it satisfies one of the conditions then the corresponding vector entry will be 1, others are zero. For a given geography, the individual vector representation for each of the occupations belonging to that geography is added to obtain the geography-level vector. Next, we cluster this vector space using a standard hierarchical spectral clustering that partitions the 21 geographies in the following way. Similar to any clustering algorithm, the number of clusters k is a parameter and is adjusted using the standard elbow method. This whole process is separately executed for all three learning algorithms TRANSE, DISTMULT, COMPGCN, and GEKC, and the clusters obtained are listed in Table 5.10. Remarkably, the clusters are broadly consistent with the popular socio-economic partitioning of the world – the *Global North* and *Global South* [148]. This observation is true for all the algorithms TRANSE, DISTMULT, COMPGCN, and GEKC despite the differences in their inner

workings. Typically, countries that are economically developed are part of the Global North, while those that are in the developing phase are considered the Global South. For example, for gender, in the case of TRANSE, the first two clusters, aka Cluster 1 and Cluster 2 (henceforth GN-1 and GN-2), are dominated by the countries from the Global North while Cluster 3 is from the Global South (henceforth GS). Likewise, gender, in the case of age too, the same division – Global North and Global South emerges from the clustering for all the KGE algorithms. The clusters for both gender and age are noted in Table 5.10; the first 3 rows denote clusters obtained in case of gender, and the last 2 rows are for age.

Quantitative evidence: To corroborate that our clusters indeed correspond to the global economic divide we further compute a set of country-level attributes, such as distance from American culture [139], GDP per capita (in USD)[18], Gini coefficient [54], Human Development Index [9], Gender Gap[19], and Individualism [140]. We report the average values of these predictors for the clusters in Table 5.11. The reported values show that, indeed, GN-1 and GN-2 have significantly different values for the above attributes compared to the GS and reveal significant differences between GN-1, GN-2, and GS, particularly in terms of the meanings conveyed by these attributes. For instance, the attribute American Cultural Distance shows that GN-1 has the lowest values, indicating that the geographies within GN-1 share very similar cultural characteristics with that of the USA. In contrast, GS exhibits much higher values, reflecting a greater cultural difference from the USA. We use average cosine similarity of the features of the countries in a cluster to measure intra-cluster similarities. The higher this similarity, the more close the countries within a cluster are.

[18]https://www.imf.org/en/Publications/WEO/weo-database/2023/April

[19]https://www.weforum.org/reports/

	TRANSE	DISTMULT	COMPGCN	GEKC
C_1:	Argentina, Australia, Canada, France, Germany, Israel, Japan, Russia, Spain, UK, USA	Australia, Canada, France, Germany, Japan, Russia, Spain, UK, USA	Australia, Spain, France, Canada, UK, Japan, Germany, Russia, USA, Brazil	Australia, Argentina, Canada, France, Germany, Israel, Japan, UK, USA
C_2:	Mexico, New Zealand, South Korea	Argentina, Brazil, Israel, Mexico, New Zealand, South Korea	Arabia, South Africa, South Korea, New Zealand, Argentina, Egypt, India, Nigeria, Israel, Mexico, Turkey	Arabia, Brazil, Egypt, Russia, India, Mexico, Nigeria, New Zealand, Spain, South Africa, South Korea, Turkey
C_3:	Arabia, Brazil, Egypt, India, Nigeria, South Africa, Turkey	Arabia, Egypt, India, Nigeria, South Africa, Turkey		
C_1:	Argentina, Canada, UK, Germany, Japan, Australia, USA, Russia, Brazil, Spain, France	Canada, Brazil, Mexico, Russia, USA, UK, France, Israel, Germany, Japan, Spain	Spain, Brazil, France, Australia, Russia, Japan, UK, Germany, Canada, USA	Australia, Argentina, Brazil, Canada, France, Germany, India, Japan, New Zealand, Spain, UK, USA
C_2:	Arabia, South Korea, Israel, Nigeria, Turkey, Mexico, India, Egypt, South Africa, New Zealand	New Zealand, South Africa, Australia, Nigeria, India, Arabia, Egypt, Turkey, Argentina, South Korea	Mexico, Argentina, New Zealand, Arabia, Turkey, Nigeria, South Africa, South Korea, India, Israel, Egypt	Arabia, Egypt, Israel, Mexico, Nigeria, Russia, South Korea, South Africa, Turkey

Table 5.10: Table showing different clusters obtained by the clustering of the features. The upper block (i.e., first 3 rows) and the lower block (i.e., last 3 rows) represent the clusters generated in the case of sensitive attributes of gender and age, respectively.

Occupations with opposite biases: To explore the differences among the clusters, we define the concept of occupations with opposite biases. An occupation is considered opposite in two clusters if it is male-biased in one cluster and female-biased in the other or vice versa. To determine whether an occupation o is opposite in two clusters c_1 and c_2, we consider four conditions:

- occupation o satisfies the condition in eq. (1) in cluster c_1 and the condition in eq. (3) in cluster c_2.
- occupation o satisfies the condition in eq. (3) in cluster c_1 and the condition in eq. (1) in cluster c_2.

- occupation o satisfies the condition in eq. (2) in c_1 and the condition in eq. (4) in cluster c_2.
- occupation o satisfies the condition in eq. (4) in cluster c_1 and the condition in eq. (2) in cluster c_2.

Table 5.12 displays the occupations following any of the four conditions mentioned above for the Global North and Global South cluster pairs. A careful inspection shows that the country-level indicators of GN-1 and GN-2, such as GDP per capita, gender gap, etc., as detailed in Table 5.11 are very close and far apart from the GS cluster. Thus, we conclude that GN-1 and GN-2 are finer distinctions of a giant Global North cluster and, hence, are combined into GN for further inspection of occupations with opposite biases. An interesting observation from the table is that male-biased occupations in the Global North and female-biased occupations in the Global South seem to be intellectually driven, such as businessperson, diplomat, artist, photographer, songwriter, etc. This possibly suggests that women tend to do jobs that are perceived as "soft" in the Global South, and men are interested in such white-collar jobs in the Global North. In contrast, female-biased occupations in the Global North and male-biased occupations in the Global South seem to be more physical activity-based, e.g., basketball players, swimmers, field hockey players, choreographers, volleyball players, athletics competitors, etc. This possibly means that women are involved in physical jobs in the Global North, while men are interested in such jobs in the Global South. This surprising finding is indeed supported by recent publications. In the article on 'Women's Employment' published in the journal *Our World in Data*, the authors studied the employment of women worldwide in the labor market [149]. They observed that the female labor force participation is highest in the richest countries (primarily in the Global North) and lowest in medium-income countries (mostly Global South). Further authors reported that in the Global North, higher levels of economic development, industrialization, and gender equality have created greater opportunities for women to engage in various sectors, including physically demanding jobs[20]. Conversely, women in the Global South frequently encounter more significant barriers to labor force participation due to conservative social norms [106], but they are

[20]https://tinyurl.com/3m8vuesx

notably present in informal sectors and less physically demanding roles. For instance, India has the highest percentage of female coding developers globally, with women comprising 22.9% of the workforce [21].

Attributes	TransE			DistMult			CompGCN		GeKC	
	GN-1	GN-2	GS	GN-1	GN-2	GS	GN	GS	GN	GS
American Cultural Distance (↓)	0.06	0.07	0.12	0.06	0.07	0.13	0.05	0.08	0.06	0.10
GDP per capita (in USD) (↑)	44,827.82	30,606	11,807	47,125	28,276	12,298	43,298	21,326	49,915.89	18,194
Gini coefficient (↓)	35.81	37.03	40.3	34.77	40.15	38.86	36.19	38.65	35.79	38.75
Human Development Index (↑)	0.91	0.85	0.71	0.88	0.74	0.70	0.89	0.78	0.92	0.78
Gender Gap Index (↑)	0.75	0.74	0.67	0.75	0.73	0.67	0.74	0.69	0.75	0.71
Individualism (↑)	65.82	42.33	43.28	69.33	44.16	44.16	66.2	44.72	70.44	43.34
Intra-cluster similarities (↑)	0.35	0.99	0.87	0.59	0.78	0.65	0.65	0.77	0.89	0.73

Table 5.11: Different country-level attributes showing social, economic, and cultural differences and intra-cluster similarities computed for the geographies grouped by clusters.

5.2.5 Summary

In our study, we examine how biases embedded in a knowledge graph, specifically Wikidata, can influence the task of link prediction, which is essential for completing a knowledge graph. Specifically, we show that the presence of social bias in terms of protected attributes, for example– gender and age in the graph, leads to certain occupations being classified as gender-inclined/age-inclined, even though they should be gender-neutral/age-neutral in an unbiased setting. We introduce a new dataset specifically

[21] https://tinyurl.com/2tpm8enm

Categories	TRANSE	DISTMULT	COMPGCN	GEKC
$(TPR_m > TPR_f)_{GN}$, $(TPR_m < TPR_f)_{GS}$	athlete, businessperson, artist, radio personality, illustrator, translator, photographer, stage actor, economist, actor, songwriter, diplomat, biologist, physician, teacher	businessperson, physician, association football player, activist, film director, radio personality, teacher, water polo player, singer-songwriter, diplomat, television producer, judoka	swimmer, tennis player, zoologist, civil servant, dancer, artist, architect, amateur wrestler, visual artist, badminton player, athletics competitor, boxer, politician, radio personality, volleyball player, rugby union player, canoeist, academic, economist, farmer, field hockey player, pianist	academic, swimmer, television presenter, athlete, long-distance runner, judge, librarian
$(TPR_m < TPR_f)_{GN}$, $(TPR_m > TPR_f)_{GS}$	lawyer, scientist, autobiographer, television presenter, designer, physician, businessperson, athletics competitor, academic, film producer, entrepreneur, fashion designer, children's writer, photographer, field hockey player, sociologist, volleyball player, swimmer, choreographer, university teacher, basketball player	musician, tennis player, zoologist, architect, university teacher, economist, politician, athletics competitor, author, autobiographer, academic	diplomat, model, tennis player, photographer, zoologist, basketball player, amateur wrestler, autobiographer, poet, volleyball player, businessperson, botanist, judoka, engineer, biologist, physician	curator, rower, physician, official
$(FPR_m > FPR_f)_{GN}$, $(FPR_m < FPR_f)_{GS}$	athletics competitor, lyricist, tennis player	businessperson, television actor, sculptor, singer-songwriter, television presenter	swimmer	-
$(FPR_m < FPR_f)_{GN}$, $(FPR_m > FPR_f)_{GS}$	singer-songwriter, film producer,	swimmer, badminton player	sculptor, stage actor, taekwondo athlete	speed skater, teacher

Table 5.12: List of occupations that belong to opposite categories in the cluster pairs– Global North (i.e., GN-1 and GN-2 together) and Global South for the attribute gender. By opposite category, we want to point out the occupations that are marked as male-biased in one cluster and female-biased in the other cluster of the cluster pairs. The pair of tuples under "Categories" in each row can be read as – the first tuple is considered for the first cluster, i.e., Global North (GN), and the second one for the other cluster, i.e., Global South (GS). For example, the first row lists the occupations that belong to the fairness category $TPR_m > TPR_f$ in GN and $TPR_m < TPR_f$ in GS.

geography	Male-biased		Female-biased		Gender-neutral	
	TRASE	DISTMULT	TRANSE	DISTMULT	TRANSE	DISTMULT
Arabia	artist, human rights activist	diplomat	singer, politician	physician	-	-
India	autobiographer, judge	-	lawyer, historian, singer-songwriter, painter	television producer, activist, television presenter	-	-
Israel	curator, sport cyclist, legal scholar	television producer, children's writer	model, illustrator, literary editor	-	-	-
Japan	videographer, physician, table tennis player, art historian	lecturer, jazz musician, tennis player, mixed martial artist	fashion model, AV idol, model, artistic gymnast	fashion designer, ice dancer, speed skater, pianost	songwriter	-
South Korea	manhwaga, athletics competitor, sport shooter, politician	comedian, fencer	pianist, musician	pianist,badminton player	-	-
Turkey	human rights activist, lawyer	economist, architect, swimmer	physician, photographer	lyricist	athletics competitor	artist
Russia	philosopher, photographer, coach	sculptor, art historian, human rights activist, painter	figure skating coach, lyricist, rower	curler, photographer, television presenter	-	-
Australia	volleyball player, farmer, editor, chess player	civil servant, trade unionist, bowls player, chess player	librarian, autobiographer, philanthropist, artistic gymnast	librarian, ballet dancer, water polo player, model	opera singer	-
NewZealand	chairperson, civil servant, film director	playwright, film actor, film director, jeweler	tennis player, illustrator, potter, sprinter	historian, tennis player	-	-
Egypt	athletics competitor, swimmer	swimmer, athletics competitor	actor, film actor	journalist, judoka	-	-
Nigeria	poet, film actor, amateur wrestler	poet	radio personality, screenwriter	rasio personality, media personality	-	-
South Africa	author, field hockey player, singer-songwriter, university teacher	author, musician, sculptor, businessperson	songwriter, water polo player, artist	water polo player, singer-songwriter, television actor	-	film actor

Table 5.13: (A): Example male-biased, female-biased, and gender-neutral occupations for each geography obtained in the training procedure by TRANSE and DISTMULT in each of the categories. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and, (5.4), (5.6), respectively.

designed for measuring social biases in knowledge graph completion tasks, sourced from Wikidata triples, incorporating two important sensitive attributes—gender and age. The dataset also includes a variety of occupations associated with human entities from 21

geography	Male-biased		Female-biased		Gender-neutral	
	TRASE	DISTMULT	TRANSE	DISTMULT	TRANSE	DISTMULT
France	pedagogue, linguist, economist, judoka	environmentalist, military person, chess plyer, businessperson	videographer, literary scholar, farmworker, environmentalist	blogger, shopkeeper, singer, performing artist	poet, swimmer	-
Germany	television producer, engineer, professor, theologian	YouTuber, herpetologist, singer songwriter	fashion designer, beach volleyball player, scenographer, alpine skier	folklorist, pornographic actor, social worker, environmentalist	volleyball player	-
Spain	dub actor, military personnel, physician, comics artist	Esperanist, political activist, canoeist	physician, comics artist, television producer, sociologist	curator, activist, fashion designer, aristocat	taekwondo athlete, television presenter	-
UK	production designer, manufacturer, podcaster, mountainer drawer	manufacturer, psychiatrist, judge, production designer	lecturer, LGBTQIA+ rights activist, rapper	landscape architect, socialite, philosopher, writer	-	-
Argentina	judge, psychologist, tennis player, translator	athelete, field hockey player, television actor	dancer, biochemist, ballet dancer, choreographer	author, economist, historian	literacy critic	film actor
Brazil	designer, YouTuber, lawyer, translator	diplomat, tennis player, civil servant	presenter, biologist, author, ichthyologist	author, historian, nurse	-	film actor
Canada	chemist, sports commentator, field hockey player, aircraft pilot	chemist, trade unionist, freestyle skier, basketball player	television presenter, designer, choreographer, stage actor	human rights activist, missionary, swimmer, television producer	artist	-
Mexico	activist, volleyball player	businessperson, sport cyclist, civil servant	academic, novelist, film producer	university teacher, fashion designer, television actor	artist	-
USA	role-playing game designer, thai boxer, real estate broker, aerospace engineer	Ameriacn football player, etcher, naval officer, architect	social worker, suffragist, prostitute, costume designer	glamour model, scholar, political adviser, pornographic actor	art historian, singer-songwriter	-

(B): Example male-biased, female-biased, and gender-neutral occupations for each geography obtained in the training procedure by TRANSE and DISTMULT in each of the categories. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and, (5.4), (5.6), respectively.

different regions globally. We quantify the biases in terms of the standard fairness metrics through our proposed noble framework– AuditLP. An interesting finding of our study

is that the extent of biases in occupations is related to the socio-economic division of the world, which separates developing and developed countries into two major groups, also known as the Global South and the Global North, respectively. Remarkably, the embedding learning algorithms, implemented from different genres – TRANSE, DISTMULT, COMPGCN and GEKC – using which the AuditLP framework draws its features from yielded the same general results, despite their differences. Our conclusions are based on the optimal hyperparameter settings for these methods, with minor adjustments to these parameters not affecting the overall observations.

geography	Gender			Age		
	male-biased	female-biased	gender-neutral	young-biased	old-biased	age-neutral
Arabia	athletics competitor, television presenter	human rights activist	_	athletics competitor,	writer	_
India	autobiographer, engineer entomologist, chess player	activist, poet, television producer, academic	_	television presenter, association football player	dancer, athletics competitor, musician	_
Israel	engineer, businessperson, television producer,	military officer, musician, judge, editor	_	model, basketball player, association football player, songwriter	singer-songwriter, singer, film director	_
Japan	lecturer, AV idol, entomologist	cross-country skier, long-distance runner, singer, weather presenter	_	field hockey player, ukiyo-e artist	entomologist, tennis player, engineer, film director	_
South Korea	film director, association football player	pianist, musician, field hockey player, songwriter	_	badminton player, manhwaga, composer, basketball player	screenwriter, taekwondo athlete	_
Turkey	television actor, diplomat, lawyer	television presenter, swimmer, model, taekwondo athlete	_	basketball player, television presenter, film producer, photographer	writer, association football manager	_
Russia	professor, artist, musicologist, fencer	speed skater, weightlifter, opera singer, sprinter	_	sport shooter, swimmer, weightlifter	economist, opera singer, politician, entrepreneur	_
Australia	canoeist, pianist, chess player	autobiographer, farmer science fiction writer, sport shooter,	painter, badminton player	artistic gymnast, basketball player, aircraft pilot	explorer, painter, television producer, academic	_
New Zealand	chairperson, basketball player, sculptor, songwriter	civil servant, rugby league player, teacher, singer	_	sport cyclist, swimmer, tennis player, singer	boxer, screenwriter	_
Egypt	_	athletics competitor, politician, swimmer	_	novelist, singer	film actor	_
Nigeria	poet, journalist, entrepreneur	radio personality, researcher, sprinter	_	boxer, businessperson	politician	_
South Africa	painter, businessperson	artist, field hockey player, writer	_	athletics competitor, badminton player, swimmer, boxer	politician, composer	_

Table 5.14: (A): Example biased occupations for each geography obtained in the training procedure by CompGCN in each of the sensitive attributes gender and age. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and (5.4), (5.6) respectively.

geography	Gender			Age		
	male-biased	female-biased	gender-neutral	young-biased	old-biased	age-neutral
France	middle-distance runner, fashion designer, violinist, athlete	pensioner, field hockey player, technician, activist	French Resistance fighter	middle-distance runner, artistic gymnast, association football player, visual artist	pensioner, anciens cadres, journalist, miscellaneous persons without work under 60 excluding retirees	_
Germany	magician, competitive diver, ski jumper, athlete	historian of the modern age, high school teacher, author, journalist	_	comics artist, ski jumper, actor, presenter	translator, association football manager, abbot, competitive diver	_
Spain	judoka, volleyball player, opera singer, religious	YouTuber, model, choreographer, political activist	_	rink hockey player, drawer, tennis player, motorcycle rider	editor, photographer, race car driver, basketball coach	_
UK	film critic, announcer, police officer, voice actor	socialite, landscape architect, music pedagogue, librarian	_	artistic gymnast, sports commentator, race car driver, professional wrestler	film editor, autobiographer, television director	_
Argentina	businessperson, biochemist	visual artist, choreographer, handball player, badminton player	_	model, basketball player, television actor, screenwriter	rower, guitarist	_
Brazil	YouTuber, ornithologist, anthropologist, sociologist	sprinter, botanist, voice actor, television presenter	_	Esperantist, rower	geographer, dancer, illustrator	_
Canada	chemist, sports commentator, short track speed skater, professional wrestler	designer, installation artist, model, writer	_	field hockey player, comics artist	historian, lawyer, photographer	_
Mexico	amateur wrestler, tennis player, taekwondo athlete, rower	academic, activist, television producer	_	professional wrestler, businessperson	swimmer, military personnel, playwright, basketball player	_
USA	war photographer, surfer, actor, musician	choir director, political adviser, writer, columnist	_	war photographer, gymnast, actor, rugby union player	bowler, merchant, writer, baseball umpire	television producer

(B): Example biased occupations for each geography obtained in the training procedure by CompGCN in each of the sensitive attributes gender and age. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and (5.4), (5.6) respectively.

geography	Young-biased		Old-biased		Age-neutral	
	TRASE	DISTMULT	TRANSE	DISTMULT	TRANSE	DISTMULT
Arabia	athletics competitor		journalist	poet	_	_
India	television presenter, association football palyer, athletics competitor	television presenter, music director	activist, businessperson,	businessperson, novelist, songwriter	songwriter	_
Israel	television presenter, singer-songwriter	television actor, translator, chess player	swimmer, lawyer	songwriter, director	dub actor	record producer
Japan	athlete, AV idol, field hockey player, tennis player	athlete, field hockey player	literary critic, film producer, sociologist	voice actor, association football manager, literary critic ,designer	coach	singer-songwriter
South Korea	baseball player	badminton player, voice actor, volleyball player	athletics competitor, voice actor, volleyball player	athletics competitor , film director	_	
Turkey	photographer		composer, author		_	
Russia	boxer, sport cyclist	volleyball player, basketball player	military personnel, painter, director	economist, middle-distance runner, lyricist, prosaist	_	
Australia	ice hockey player, competitive diver, opera singer	judoka, television producer, sport cyclist, figure skater	trade unionist, explorer, judoka	explorer, businessperson, civil servant	_	
New Zealand	singer, writer, rower rugby league player	academic, writer	musician, sport cyclist, cricketer	screenwriter, tennis player	_	
Egypt	singer, screenwriter	film director	film director, politician, composer	composer, film actor	_	
Nigeria	athletics competitor	writer, athletics competitor, boxer	businessperson	journalist, businessperson	_	
South Africa	photographer, boxer	boxer, film director, badminton player	swimmer, athletics competitor	businessperson, composer	_	

Table 5.15: (A): Example young-biased, old-biased, and age-neutral occupations for each geography obtained in the training procedure by TRANSE and DISTMULT in each of the categories. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and (5.4), (5.6) respectively.

geography	Young-biased		Old-biased		Age-neutral	
	TRASE	DISTMULT	TRANSE	DISTMULT	TRANSE	DISTMULT
France	ice hockey player, volleyball player, veterinarian, colonial administrator	association football referee, activist, association football player, amateur wrestler	cinematographer, business executive chauffer, Officer of the French Navy	nun, manual worker ,writer, businessperson	researcher	
Germany	disc jockey, farmer, comics artist, swimmer	luger, ruler, ski jumper	rugby union player, poet lawyer, police officer	water polo player, merchant, theologian, linguist	-	
Spain	tennis player, filmmaker, mountaineer, rink hockey player	rink hockey player, sport cyclist, trade unionist, tennis player	editor, sculptor, aristocrat, sailor	theatrical director, sculptor, linguist, drawer	film director	
UK	athlete, alpine skier, darts player, film maker	Formula One driver, police officer, association football manager, racing automobile driver	theatrical director, sports commentator, figure skater, literary critic	computer scientist, graphic designer, soldier	-	-
Argentina	university teacher, musician	television presenter, teacher, racing automobile driver, basketball player	film director, screenwriter, stage actor, playwright	businessperson, screenwriter, alpine skier, university teacher	-	
Brazil	engineer, official, association football manager, dancer	racing automobile driver, police officer, association football manager	architect, painter, artist, farmer	Esperantist, vocalist	vocalist	
Canada	tennis player, baseball player, cinematographer, film editor	comics artist, engineer, lacrosse player, figure skater	cricketer, historian, theatrical director, comics artist	sprinter, playwright, visual artist		
Mexico	basketball player, baseball player, film producer	basketball player, activist, singer, songwriter	painter, activist, playwright	engineer, businessperson, television producer, stage actor	-	poet
USA	war photographer, rapper, showrunner, chief executive officer	war photographer, surfer, circus performer, singer	educator, political scientist, writer, librarian, choir director, slaveholder	writer, chess player	basketball player	pianist

(B): Example young-biased, old-biased, and age-neutral occupations for each geography obtained in the training procedure by TRANSE and DISTMULT in each of the categories. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and (5.4), (5.6) respectively.

geography	Gender			Age		
	male-biased	female-biased	gender-neutral	young-biased	old-biased	age-neutral
Arabia	human rights activist, politician	diplomat, artist		military personnel, writer	politician, singer	
India	diplomat, human rights activist, chess player	judge, lawyer, fashion designer, artist		athletics competitor, association football player, amateur wrestler, composer	playback singer, songwriter, lyricist	
Israel	businessperson, television presenter, opinion journalist	fashion designer, model, childrenś writer, dancer		film producer, journalist, composer, songwriter	mathematician, military personnel	
Japan	military personnel, oracle, coach, graphic designer	artistic gymnast, childrenś writer, playwright, handball player		model, basketball player, broadcast writer, researcher	choreographer, musicologist, rower, artist	
South Korea	dancer, basketball player, short track speed skater, comedian	fencer, swimmer, screenwriter, rower	poet, association football player	volleyball player, composer	politician, film director, amateur wrestler	
Turkey	translator, photographer	lyricist, academic, archaeologist, musician	actor		composer, film director	
Russia	record producer, paleontologist, philosopher, artist	blogger, athlete, stage actor, lyricist	visual artist	radio personality, volleyball player, public figure	teacher, military personnel, figure skater, artist	
Australia	trade unionist, athlete	artistic gymnast, sailor, model, water polo player		field hockey player, civil servant, racing driver, judoka	artist, singer-songwriter, sprinter	rugby league player
New Zealand	musician, trade unionist, botanical collector, political candidate	librarian, chemist, swimmer, field hockey player	politician	tennis player, military personnel, athlete, cricketer	swimmer, actor	
Egypt	politician, translator	singer		composer	journalist, politician	
Nigeria	journalist, presenter, sprinter, athletics competitor	businessperson, film producer, association football player		businessperson, association football player, athletics competitor	politician, artist	
South Africa	businessperson, painter	botanist, film director	squash player	cricketer	artist, boxer	

Table 5.16: (A): Example biased occupations for each geography obtained in the training procedure by GEKC in each of the sensitive attributes gender and age. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and (5.4), (5.6), respectively.

geography	Gender			Age		
	male-biased	female-biased	gender-neutral	young-biased	old-biased	age-neutral
France	architect, radio personality, comics artist	entrepreneur, explorer, employee	university teacher, association football player	bobsledder, rugby league player, flying ace, music pedagogue	penciller, anciens cadres, cadres de la fonction publique, weightlifter	
Germany	choir director, engineer, science fiction writer, art dealer	nurse, German scholar, javelin thrower, speed skater		judoka, sport shooter, coach	artistic gymnast, German scholar, costume designer, cross-country skier	table tennis player, badminton player
Spain	anarcho-syndicalist, zoologist, architect	long-distance runner, pornographic actor, blogger, archivist	association football player, writer	explorer, model, humorist, water polo player	art historian, diplomat, university teacher, basketball coach	
UK	museum director, SOE agent, engineer, head teacher	lyricist, dancer, curator, watercolorist		sports commentator, socialite, beauty pageant contestant, sprinter	entertainer, water polo player, fencer	athletics competitor
Argentina	scientist, stage actor	visual artist, architect, teacher, association football player		trade unionist, racing automobile driver, sport cyclist, swimmer	composer, screenwriter, botanist, association football manager	
Brazil	illustrator, official, singer-songwriter, herpetologist	athlete, dancer, YouTuber, artist	zoologist	farmer, high school teacher	painter, record producer, official	television actor, actor
Canada	civil servant, librarian, boxer, canoeist	violinist, videographer, recording artist, ecologist		blogger, filmmaker, cricketer	film editor, author, writer	musician, judoka
Mexico	taekwondo athlete, businessperson	teacher, zoologist, dancer, philosopher	association football player	basketball player, badminton player, artist	lawyer, activist, painter, boxer	television actor
USA	surgeon, American football player, music video director, sport shooter	musical theatre actor, folklorist, civil rights advocate, real estate broker	badminton player	pornographic actor, triathlete, mandolinist, kickboxer	film critic, political activist, civil rights advocate, trumpeter	

(B): Example biased occupations for each geography obtained in the training procedure by GEKC in each of the sensitive attributes gender and age. Here, red-colored and blue-colored occupations denote the occupations satisfying equations (5.3), (5.5) and (5.4), (5.6), respectively.

Chapter 6

Conclusion and Future Work

In this concluding chapter we summarize the contributions of this thesis, point out the limitations and suggest possible future directions.

6.1 Summary of contributions

This thesis seeks to understand the challenges of enriching open-sourced content on collaborative platforms, specifically Wikipedia and Wikidata. In this section, we summarize the key contributions made in this thesis.

6.1.1 Ecosystem of quality in Wikipedia

We aim to address the problem of manual quality assessment in English Wikipedia articles. First, we studied the evolution of a Wikipedia articles with respect to the article quality ranking – FA, A, GA, B, C, Start, Stub – ordered in terms of decreasing quality, and identified various non-intuitive patterns that demonstrate the nuances of this ecosystem. Next, we developed an automated data-driven approach to detect early signals

influencing the quality change of articles. We framed this as a change point detection problem, representing an article as a time series of consecutive revisions and encoding each revision with a set of intuitive features. Our proposed method is possibly the first to offer a novel unsupervised page-level approach for detecting dynamic quality switches, which can assist in automatic content monitoring on Wikipedia.

6.1.2 Ecosystem of knowledge equity in Wikipedia

Here we aimed to address the content disparity in Wikipedia articles between high-resourced and low-resourced language versions. We chose English as the high-resourced language and Hindi as the relatively low-resourced one for our experiments. We develop a framework for effectively transferring knowledge from enriched English articles to their less rich Hindi counterparts on the same topics, leveraging the machine-translation capabilities of the SOTA language model –`IndicTrans`. Further, the framework includes a sub-module to adapt knowledge from external resources and align them with Wikipedia's NPOV standards to be integrated with the Hindi articles.

6.1.3 Ecosystem of bias and fairness in Wikidata

Auditing data and algorithm bias

We presented experiments auditing the societal biases embedded in Wikidata, specifically data and algorithmic biases that surface when knowledge triples are converted into embeddings by embedding learning algorithms. We examined how two key design choices that influence bias measurement approaches: (i) the selection of geo-social data, i.e., Wikidata triples related to different geographical regions, and (ii) the choice of knowledge graph embedding learning algorithms, such as TRANSE and COMPLEX. Our findings reveal the variability of biases across these two design choices, challenging the popular notion of "one-size-fits-all."

Auditing bias and fairness in link prediction in Wikidata

We explore the interplay between bias and fairness in the context of the downstream link prediction task within the Wikidata ecosystem. Our analysis consists of two steps. First, we curated a large dataset of 3.2 million human entities from 21 different geographies across the globe. Next, we introduced a framework named AUDITLP, which analyzes unfairness based on sensitive attributes, such as gender and age, in predicting the occupations of human entities in our dataset. Further, we examined the impact of two critical design choices – geosocial data selection and embedding learning algorithms on fairness in LP. Our micro-level analysis reveals numerous biases, ultimately resulting in a clear partition of the world into two distinct regions: the Global North and the Global South, characterized by their different geo-social and economic attributes.

6.2 Limitations

- In the third chapter, we discussed several offline multivariate parametric change point detection algorithms. While offline algorithms typically offer greater accuracy in identifying change points, online algorithms are more suited for real-time applications, as they can detect changes immediately as new data arrives. In the context of monitoring quality changes in Wikipedia articles, online algorithms can predict whether a recent revision signals a potential quality change, making them ideal for real-time monitoring. Our proposed pipeline can be implemented for real-time use by incorporating online change point algorithms.

- In the fourth chapter, we developed a framework to transfer knowledge from high-resource languages, like English, to low-resource languages, such as Hindi, by leveraging the generative capabilities of advanced machine translation systems. However, this process can inadvertently exacerbate cultural biases across the language versions. As previous research [37] has highlighted, cultural bias often results in the over-representation of topics relevant to a specific context, say Western, while under-representing or misrepresenting those from non-Western regions. For

example, English Wikipedia articles tend to provide more comprehensive coverage of topics related to North America and Europe, while key topics pertaining to Africa, Asia, and the Global South are less frequently detailed or absent. When translating content from one language version to another, specifically from English to Hindi, such biases and stereotypes of resource-enriched languages can easily be carried over to the low-resource language. While our automated approach helps bridge the information gap in low-resource languages, it might also increase the risk of overshadowing the subtle cultural elements that are vital to the local context. To mitigate this, language-specific domain expert editors would need to conduct thorough manual reviews before integrating the generated content.

- In the first study of the fifth chapter, where we audited biases embedded in Wikidata, our experiments were limited to two knowledge graph embedding algorithms: TRANSE and COMPLEX. These algorithms respectively represent the translational and decomposition approaches for learning knowledge graph embeddings. Including other neural network based approaches can further strengthen the observations.

- In the second study of the fifth chapter, we employed two key fairness metrics —- *equalized odds* and *equal opportunity* to assess whether an occupation is biased. Our approach can be extended to classify occupations by incorporating additional ML fairness definitions, such as demographic parity and predictive parity. So far, in our experiments, we focused on two sensitive attributes: gender and age. Our proposed framework can also be adapted to explore other sensitive attributes, such as ethnicity, religion, etc., though extracting these attributes from Wikidata triples may present extreme challenges.

6.3 Future research directions

In this section, we discuss some of the future directions that have been opened up by this thesis.

6.3.1 Early detection of quality in multiple language versions of Wikipedia

In the first chapter, we analyzed the evolution of article quality in our sample dataset of English Wikipedia articles. In addition, we proposed an unsupervised page-level approach to detect quality change points for individual articles without the extra cost of model training. We framed this as a change point detection problem and implemented offline change point detection algorithms. In the future, one can aim to deploy an early quality change monitoring system that integrates online and offline change point detection algorithms. Our framework can also be adapted for other language versions, particularly low-resourced ones that lack a quality class hierarchy, to monitor quality changes. A potential improvement could be the addition of language-agnostic features to support the early detection of quality changes in multiple low-resourced languages.

6.3.2 Inclusion of knowledge resources and verifiable of knowledge equity

In the second chapter, we aimed to bridge the information gap between high-resourced languages (i.e., English) and low-resourced languages (i.e., Hindi). While transferring knowledge from high-resourced to low-resourced languages, we augmented content from external resources to enrich knowledge further. For the collection of Wikipedia articles in our dataset that belong to the person category, we have utilized biographical writings as external resources. As immediate future work, our framework could be applied to Wikipedia articles in different categories, such as politics, arts, science, medicine, etc., and additional knowledge could be gathered from large collections of digital news forums related to these particular topics. The SOTA RAG framework can be trained on this knowledge corpus to extract relevant and factual information related to the given section content. Further, there is scope to introduce qualitative metrics that can evaluate the quality of transferred knowledge in low-resource languages directly. Another potential future direction could involve incorporating factually correct references in typical low-resource languages by analyzing

the newly added content, thus supporting Wikipedia's prime concern of verifiable content.

6.3.3 Further audit and mitigation of social biases in Wikidata

In the third chapter, we focused on surveying social biases in Wikidata and further their propagation to the very crucial downstream application – link prediction. In the first work, we observed the biases in the case of the sensitive attribute – gender. As a future work, one can extend our proposed pipeline to multiple sensitive attributes, such as a combination of gender and race, by modifying the bias score computation function and further investigating fine-grained social biases across the design choices outlined in our work. One important issue, however, is to develop an effective and accurate way to identify the race of the person entities from Wikidata. Further, one can invent suitable approaches to mitigate such biases where multiple sensitive attributes play an important role.

State-of-the-art bias mitigation algorithms leverage techniques like adversarial learning [17] and embedding fine-tuning [64] to reduce biases in knowledge graph embeddings. However, current approaches primarily focus on entity-wise bias within knowledge graphs, often overlooking implicit fairness issues in more complex, higher-order relationships. For instance, while entity-wise debiasing methods may address biases between entities like `people` and `gender` by making `gender` indistinguishable from `people`, they fail to account for relational biases in multi-hop connections, such as the bias between `occupation` and `gender` in the triple `occupation--people--gender`. In our audit of biases embedded in Wikidata, we specifically have targeted multi-hop relational bias, focusing on how `gender` influences the prediction of `occupation`. This underscores the need for approaches that address multi-hop relational bias. Recent work by Chuang et al. [44] introduces methods to mitigate multi-hop relational biases while preserving the proximity information between entities and relations in knowledge graphs. Immediate research in this direction should explore how varying knowledge graphs across geographies might affect prediction outcomes of debiased embedding generated by advanced bias mitigation strategies. In the case of our second work on link prediction, a promising future direction would be to employ our proposed framework

AUDITLP to other knowledge graph data, such as Yahoo, DBpedia, etc, for investigating fairness in these knowledge graphs.

Bibliography

[1] Sayantan Adak, Pauras Mangesh Meher, Paramita Das, and Animesh Mukherjee. Reversum: A multi-staged retrieval-augmented generation method to enhance wikipedia tail biographies through personal narratives. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 732–750, 2025.

[2] Eytan Adar, Michael Skinner, and Daniel S Weld. Information arbitrage across multi-lingual wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103, 2009.

[3] Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of NAACL*, pages 3554–3565, 2021.

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

[5] AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[6] Desislava Aleksandrova, François Lareau, and Pierre André Ménard. Multilingual sentence-level bias detection in wikipedia. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 42–51, 2019.

[7] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8825–8845, 2021.

[8] Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[9] Sudhir Anand and Amartya Sen. Human development index: Methodology and measurement. 1994.

[10] Maik Anderka, Benno Stein, and Nedim Lipka. Predicting quality flaws in user-generated content: the case of wikipedia. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 981–990, 2012.

[11] Ofer Arazy and Oded Nov. Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 233–236, 2010.

[12] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. Functional roles and career paths in wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1092–1105, 2015.

[13] Ofer Arazy, Johannes Daxenberger, Hila Lifshitz-Assaf, Oded Nov, and Iryna Gurevych. Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge coproduction. *Information Systems Research*, 27(4): 792–812, 2016.

[14] Ofer Arazy, Hila Liifshitz-Assaf, Oded Nov, Johannes Daxenberger, Martina Balestra, and Coye Cheshire. On the" how" and" why" of emergent role behaviors in wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2039–2051, 2017.

[15] Ofer Arazy, Aron Lindberg, Shakked Lev, Kexian Wu, and Alex Yarovoy. Emergent routines in peer-production: Examining the temporal evolution of wikipedia's work sequences. *ACM Transactions on Social Computing*, 3(1):1–24, 2020.

[16] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[17] Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. Adversarial learning for debiasing knowledge graph embeddings. *arXiv preprint arXiv:2006.16309*, 2020.

[18] Akhil Arora, Martin Gerlach, Tiziano Piccardi, Alberto García-Durán, and Robert West. Wikipedia reader navigation: When synthetic data is enough. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 16–26, 2022.

[19] Vevake Balaraman, Simon Razniewski, and Werner Nutt. Recoin: relative completeness in wikidata. In *Companion Proceedings of the The Web Conference 2018*, pages 1787–1792, 2018.

[20] Ivana Balažević, Carl Allen, and Timothy M Hospedales. Hypernetwork knowledge graph embeddings. In *International Conference on Artificial Neural Networks*, pages 553–565. Springer, 2019.

[21] Martina Balestra, Lior Zalmanson, Coye Cheshire, Ofer Arazy, and Oded Nov. It was fun, but did it last? the dynamic interplay between fun motives and contributors' activity in peer production. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–13, 2017.

[22] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[23] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084, 2012.

[24] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[25] Elias Bassani and Marco Viviani. Automatically assessing the quality of wikipedia contents. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 804–807, 2019.

[26] Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K. Singh. Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):43–54, May 2022. doi: 10.1609/icwsm.v16i1.19271.

[27] Taryn Bipat, David W McDonald, and Mark Zachry. Do we all talk before we type? understanding collaboration in wikipedia language editions. In *Proceedings of the 14th International Symposium on Open Collaboration*, pages 1–11, 2018.

[28] Joshua E Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, pages 1095–1096, 2008.

[29] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[30] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[31] Avishek Bose and William Hamilton. Compositional fairness constraints for

graph embeddings. In *International Conference on Machine Learning*, pages 715–724. PMLR, 2019.

[32] Gosse Bouma, Sergio Duarte, and Zahurul Islam. Cross-lingual alignment and completion of wikipedia templates. In *Proceedings of the third international workshop on cross lingual information access: Addressing the information need of multilingual societies (CLIAWS3)*, pages 21–29, 2009.

[33] Styliani Bourli and Evaggelia Pitoura. Bias in knowledge graph embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 6–10. IEEE, 2020.

[34] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[35] Moira Burke and Robert Kraut. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 27–36, 2008.

[36] Benjamin Cabrera, Björn Ross, Marielle Dado, and Maritta Heisel. The gender gap in wikipedia talk pages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[37] Ewa S Callahan and Susan C Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

[38] Meng Cao, Jianqing Song, Jinliang Yuan, Baoming Zhang, and Chongjun Wang. Fairhelp: Fairness-aware heterogeneous information network embedding for link prediction. In *International Conference on Database Systems for Advanced Applications*, pages 320–330. Springer, 2023.

[39] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM*

joint meeting on european software engineering conference and symposium on the foundations of software engineering, pages 429–440, 2021.

[40] Kaylea Champion, Nora McDonald, Stephanie Bankes, Joseph Zhang, Rachel Greenstadt, Andrea Forte, and Benjamin Mako Hill. A forensic qualitative analysis of contributions to wikipedia from anonymity seeking users. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[41] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[42] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

[43] Boreum Choi, Kira Alexander, Robert E Kraut, and John M Levine. Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 107–116, 2010.

[44] Yu-Neng Chuang, Kwei-Herng Lai, Ruixiang Tang, Mengnan Du, Chia-Yuan Chang, Na Zou, and Xia Hu. Fair-rgnn: Mitigating relational bias on knowledge graphs. *ACM Transactions on Knowledge Discovery from Data*, 2024.

[45] Benjamin Collier and Julia Bear. 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*, pages 383–392, 2012.

[46] Michele Coscia and Michael Szell. Multilayer graph association rules for link prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 129–139, 2021.

[47] Yuanfei Dai, Shiping Wang, Neal N Xiong, and Wenzhong Guo. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics*, 9(5):750, 2020.

[48] Quang-Vinh Dang and Claudia-Lavinia Ignat. Measuring quality of collaboratively edited documents: The case of wikipedia. In *2016 IEEE 2nd international conference on collaboration and internet computing (CIC)*, pages 266–275. IEEE, 2016.

[49] Quang Vinh Dang and Claudia-Lavinia Ignat. Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 27–30, 2016.

[50] Quang-Vinh Dang and Claudia-Lavinia Ignat. An end-to-end learning solution for assessing the quality of wikipedia articles. In *Proceedings of the 13th International Symposium on Open Collaboration*, pages 1–10, 2017.

[51] David Danks and Alex John London. Algorithmic bias in autonomous systems. In *Ijcai*, volume 17, pages 4691–4697, 2017.

[52] Paramita Das, Isaac Johnson, Diego Saez-Trumper, and Pablo Aragón. Language-agnostic modeling of wikipedia articles for content quality assessment across languages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1924–1934, 2024.

[53] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. Measuring article quality in wikipedia using the collaboration network. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 464–471. IEEE, 2015.

[54] Fernando G De Maio. Income inequality measures. *Journal of Epidemiology & Community Health*, 61(10):849–852, 2007.

[55] Gianluca Demartini. Implicit bias in crowdsourced knowledge graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 624–630, 2019.

[56] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[57] Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building socio-culturally inclusive stereotype resources with community engagement. *Advances in Neural Information Processing Systems*, 36, 2024.

[58] Virginia Dignum. Relational artificial intelligence. *arXiv preprint arXiv:2202.07446*, 2022.

[59] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[60] Fahim Faisal and Antonios Anastasopoulos. Geographic and geopolitical biases of language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163, 2023.

[61] Yiwei Fan and Xiaoling Lu. An online bayesian approach to change-point detection for categorical data. *Knowledge-Based Systems*, page 105792, 2020.

[62] Elena Filatova. Directions for exploiting asymmetries in multilingual wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)*, pages 30–37, 2009.

[63] Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. Measuring social bias in knowledge graph embeddings. *arXiv preprint arXiv:1912.02761*, 2019.

[64] Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7332–7345, 2020.

[65] Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. Measuring social bias in knowledge graph embeddings, 2020.

[66] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[67] Andrea Forte, Nazanin Andalibi, and Rachel Greenstadt. Privacy, anonymity, and perceived risk in open collaboration: A study of tor users and wikipedians. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1800–1811, 2017.

[68] Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*, 2023.

[69] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

[70] R Stuart Geiger and Aaron Halfaker. Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 861–870, 2013.

[71] R Stuart Geiger and Aaron Halfaker. Operationalizing conflict and cooperation between automated software agents in wikipedia: A replication and expansion of'even good bots fight'. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–33, 2017.

[72] José Gustavo Góngora-Goloubintseff. The falklands/malvinas war taken to the wikipedia realm: a multimodal discourse analysis of cross-lingual violations of the neutral point of view. *Palgrave Communications*, 6(1):1–9, 2020.

[73] Bhanu Prakash Reddy Guda, Sasi Bhushan Seelaboyina, Soumya Sarkar, and Animesh Mukherjee. Nwqm: A neural quality assessment framework for

wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8396–8406, 2020.

[74] Scott A Hale. Multilinguals and wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, pages 99–108, 2014.

[75] Aaron Halfaker and R Stuart Geiger. Ores: Lowering barriers with participatory machine learning in wikipedia. *arXiv preprint arXiv:1909.05189*, 2019.

[76] Aaron Halfaker, Aniket Kittur, and John Riedl. Don't bite the newbies: how reverts affect the quantity and quality of wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*, pages 163–172, 2011.

[77] Felix Hausdorff. *Mengenlehre*. Walter de Gruyter Berlin, 1927.

[78] Kaylea Haynes, Paul Fearnhead, and Idris A Eckley. A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293–1305, 2017.

[79] Shiqing He, Allen Yilun Lin, Eytan Adar, and Brent Hecht. The_tower_of_babel. jpg: diversity of visual encyclopedic knowledge across wikipedia language editions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[80] Brent Hecht and Darren Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies*, pages 11–20, 2009.

[81] Brent Hecht and Darren Gergle. The tower of babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 291–300, 2010.

[82] Livnat Herzig, Alex Nunes, and Batia Snir. An annotation scheme for automated bias detection in wikipedia. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 47–55, 2011.

[83] Laura Hollink, Astrid Van Aggelen, and Jacco Van Ossenbruggen. Using the web of data to study gender differences in online knowledge sources: the case of the european parliament. In *Proceedings of the 10th ACM Conference on Web Science*, pages 381–385, 2018.

[84] Gary Hsieh, Youyang Hou, Ian Chen, and Khai N Truong. " welcome!" social and psychological predictors of volunteer socializers in online communities. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 827–838, 2013.

[85] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 243–252, 2007.

[86] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 105–113, 2019.

[87] Christoph Hube and Besnik Fetahu. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786, 2018.

[88] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[89] Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24, 2018.

[90] Corey Brian Jackson, Kevin Crowston, and Carsten Østerlund. Did they login? patterns of anonymous contributions in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–16, 2018.

[91] Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. Debiasing knowledge graphs: Why female presidents are not like female popes. In *International Semantic Web Conference (P&D/Industry/BlueSky)*, 2018.

[92] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696, 2015.

[93] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

[94] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019.

[95] Mengyin Jiang, Shirley KM Wong, Harry KS Chung, Yang Sun, Janet H Hsiao, Jie Sui, and Glyn W Humphreys. Cultural orientation of self-bias in perceptual matching. *Frontiers in Psychology*, 10:1469, 2019.

[96] Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. Global gender differences in wikipedia readership. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 254–265, 2021.

[97] David Jurgens and Tsai-Ching Lu. Temporal motifs reveal the dynamics of editor interactions in wikipedia. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[98] Gerald C Kane, Jeremiah Johnson, and Ann Majchrzak. Emergent life cycle: The tension between knowledge change and knowledge retention in open online coproduction communities. *Management Science*, 60(12):3026–3048, 2014.

[99] Daphna Keidar, Mian Zhong, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. Towards automatic bias detection in knowledge graphs, 2021.

[100] Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. Spatio-temporal analysis of reverted wikipedia edits. In *ICWSM*, pages 122–131, 2017.

[101] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[102] Suin Kim, Sungjoon Park, Scott A Hale, Sooyoung Kim, Jeongmin Byun, and Alice H Oh. Understanding editing behaviors in multilingual wikipedia. *PloS one*, 11(5):e0155305, 2016.

[103] Aniket Kittur and Robert E Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46, 2008.

[104] Aniket Kittur and Robert E Kraut. Beyond wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224, 2010.

[105] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, 2007.

[106] Stephan Klasen. What explains uneven female labor force participation levels and trends in developing countries? *The World Bank Research Observer*, 34(2): 161–197, 2019.

[107] Maximilian Klein, Thomas Maillart, and John Chuang. The virtuous circle of wikipedia: Recursive measures of collaboration structures. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1106–1115, 2015.

[108] Piotr Konieczny and Maximilian Klein. Gender gap through time and space: A journey through wikipedia biographies via the wikidata human gender indicator. *New Media & Society*, 20(12):4608–4633, 2018.

[109] Piotr Konieczny and Włodzimierz Lewoniewski. Quantifying americanization: Coverage of american topics in different wikipedias. *Social Science Computer Review*, page 08944393231220165, 2024.

[110] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

[111] Angelie Kraft and Ricardo Usbeck. The lifecycle of "facts": A survey of social bias in knowledge graphs. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 639–652, 2022.

[112] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289, 2020.

[113] Cliff Lampe, Jonathan Obar, Elif Ozkaya, Paul Zube, and Alcides Velasquez. Classroom wikipedia participation effects on future intentions to contribute. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 403–406, 2012.

[114] Marc Lavielle and Gilles Teyssiere. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

[115] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. Why the world reads wikipedia: Beyond english speakers. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 618–626, 2019.

[116] Włodzimierz Lewoniewski. Enrichment of information in multilingual wikipedia based on quality analysis. In *Business Information Systems Workshops: BIS 2017 International Workshops, Poznań, Poland, June 28-30, 2017, Revised Papers 20*, pages 216–227. Springer, 2017.

[117] Włodzimierz Lewoniewski. Measures for quality assessment of articles and infoboxes in multilingual wikipedia. In *Business Information Systems Workshops: BIS 2018 International Workshops, Berlin, Germany, July 18–20, 2018, Revised Papers 21*, pages 619–633. Springer, 2019.

[118] Wlodzimierz Lewoniewski, Krzysztof Wecel, and Witold Abramowicz. Quality and importance of wikipedia articles in different languages. In *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings 22*, pages 613–624. Springer, 2016.

[119] Wlodzimierz Lewoniewski, Krzysztof Wecel, and Witold Abramowicz. Relative quality and popularity evaluation of multilingual wikipedia articles. In *Informatics*, volume 4, page 43. MDPI, 2017.

[120] Wlodzimierz Lewoniewski, Krzysztof Wecel, and Witold Abramowicz. Multilingual ranking of wikipedia articles with quality and popularity assessment in different topics. *Computers*, 8(3):60, 2019.

[121] Ang Li, Zheng Yao, Diyi Yang, Chinmay Kulkarni, Rosta Farzan, and Robert E Kraut. Successful online socialization: Lessons from the wikipedia education program. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–24, 2020.

[122] Muyan Li, Heshen Zhou, Jingrui Hou, Ping Wang, and Erpei Gao. Is cross-linguistic advert flaw detection in wikipedia feasible? a multilingual-bert-based transfer learning approach. *Knowledge-Based Systems*, 252:109330, 2022.

[123] Xinyi Li, Jintao Tang, Ting Wang, Zhunchen Luo, and Maarten De Rijke. Automatically assessing wikipedia article quality by exploiting article–editor networks. In *European Conference on Information Retrieval*, pages 574–580. Springer, 2015.

[124] Yanying Li, Xiuling Wang, Yue Ning, and Hui Wang. Fairlp: Towards fair link prediction on social network graphs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 628–639, 2022.

[125] Nedim Lipka and Benno Stein. Identifying featured articles in wikipedia: writing style matters. In *Proceedings of the 19th international conference on World wide web*, pages 1147–1148, 2010.

[126] Jun Liu and Sudha Ram. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)*, 2(2):1–23, 2011.

[127] Lorenzo Loconte, Nicola Di Mauro, Robert Peharz, and Antonio Vergari. How to turn your knowledge graph embeddings into generative models. *Advances in Neural Information Processing Systems*, 36, 2024.

[128] Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, 2019.

[129] Karthic Madanagopal and James Caverlee. Towards ongoing detection of linguistic bias on wikipedia. In *Companion Proceedings of the Web Conference 2021*, pages 629–631, 2021.

[130] Karthic Madanagopal and James Caverlee. Improving linguistic bias detection in wikipedia using cross-domain adaptive pre-training. In *Companion Proceedings of the Web Conference 2022*, pages 1301–1309, 2022.

[131] S. Maity, Jot Sarup Singh Sahni, and Animesh Mukherjee. Analysis and prediction of question topic popularity in community q&a sites: A case study of quora. In *ICWSM*, 2015.

[132] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7):655–665, 2016.

[133] Rohin Manvi, Samar Khanna, Marshall Burke, David B Lobell, and Stefano

Ermon. Large language models are geographically biased. In *Forty-first International Conference on Machine Learning*.

[134] David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

[135] Marc Miquel-Ribé and David Laniado. Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in physics*, 6:54, 2018.

[136] Marc Miquel-Ribé and David Laniado. The wikipedia diversity observatory: A project to identify and bridge content gaps in wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–4, 2020.

[137] Jonathan T Morgan and Anna Filippova. 'welcome'changes? descriptive and injunctive norms in a wikipedia sub-community. 2018.

[138] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 839–848, 2013.

[139] Michael Muthukrishna, Adrian V Bell, Joseph Henrich, Cameron M Curtin, Alexander Gedranovich, Jason McInerney, and Braden Thue. Beyond western, educated, industrial, rich, and democratic (weird) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science*, 31(6):678–701, 2020.

[140] Ivana Načinović Braje, Maja Klindžić, and Lovorka Galetić. The role of individual variable pay in a collectivistic culture society: An evaluation. *Economic research-Ekonomska istraživanja*, 32(1):1352–1372, 2019.

[141] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

[142] Sneha Narayan, Jake Orlowitz, Jonathan Morgan, Benjamin Mako Hill, and Aaron Shaw. 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*, pages 1785–1799, 2017.

[143] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 327–333, 2018.

[144] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Icml*, 2011.

[145] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[146] Safiya Umoja Noble. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press, 2018.

[147] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[148] Lemuel Ekedegwa Odeh. A comparative analysis of global north and global south economies. 2010.

[149] Esteban Ortiz-Ospina, Sandra Tzvetkova, and Max Roser. Women's employment. *Our world in data*, 2024.

[150] ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[151] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 51–60, 2009.

[152] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[153] Monica Lestari Paramita, Paul Clough, and Robert Gaizauskas. Using section headings to compute cross-lingual similarity of wikipedia articles. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39*, pages 633–639. Springer, 2017.

[154] Sungjoon Park, Suin Kim, Scott Hale, Sooyoung Kim, Jeongmin Byun, and Alice Oh. Multilingualwikipedia: Editors of primary language contribute to more complex articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 83–88, 2015.

[155] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.

[156] Matthew E. Peters, Mark Neumann, Robert L Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *EMNLP*, 2019.

[157] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[158] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489, 2020.

[159] Millenio Ramadizsa, Fariz Darari, Werner Nutt, and Simon Razniewski. Knowledge gap discovery: A case study of wikidata. In *The 4th Wikidata Workshop*, 2023.

[160] Narun Raman, Nathaniel Sauerberg, Jonah Fisher, and Sneha Narayan. Classifying wikipedia article quality with revision history networks. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–7, 2020.

[161] Reinhard Rapp, Serge Sharoff, and Bogdan Babych. Identifying word translations from comparable documents without a seed lexicon. In *LREC*, pages 460–466, 2012.

[162] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1650–1659, 2013.

[163] Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv preprint arXiv:2008.12314*, 2020.

[164] Clemens Reimann, Peter Filzmoser, and Robert G. Garrett. Background and threshold: critical comparison of methods of determination. *Science of The Total Environment*, 346(1):1–16, 2005. doi: https://doi.org/10.1016/j.scitotenv.2004.11.023.

[165] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15 (2):1–49, 2021.

[166] Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. A topic-aligned multilingual corpus of wikipedia articles for studying information asymmetry in low resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2373–2380, 2020.

[167] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[168] Anna Samoilenko, Florian Lemmerich, Katrin Weller, Maria Zens, and Markus Strohmaier. Analysing timelines of national histories across wikipedia editions: A comparative computational approach. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 210–219, 2017.

[169] Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216, 2009.

[170] Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4498–4507, 2020.

[171] Andrew Jhon Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.

[172] Zaina Shaik, Filip Ilievski, and Fred Morstatter. Analyzing race and country of citizenship bias in wikidata. *arXiv preprint arXiv:2108.05412*, 2021.

[173] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3060–3067, 2019.

[174] Shilpy Sharma, David A Swayne, and Charlie Obimbo. Trend analysis and change point techniques: a survey. *Energy, Ecology and Environment*, 1(3): 123–130, 2016.

[175] Aili Shen, Jianzhong Qi, and Timothy Baldwin. A hybrid model for quality assessment of wikipedia articles. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 43–52, 2017.

[176] Aili Shen, Bahar Salehi, Timothy Baldwin, and Jianzhong Qi. A joint model for

multimodal document quality assessment. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 107–110. IEEE, 2019.

[177] Harry Shomer, Wei Jin, Wentao Wang, and Jiliang Tang. Toward degree bias in embedding-based knowledge graph completion. In *Proceedings of the ACM Web Conference 2023*, pages 705–715, 2023.

[178] Chinthani Sugandhika and Supunmali Ahangama. Assessing information quality of wikipedia articles through google's eat model. *IEEE Access*, 10:52196–52209, 2022.

[179] Róbert Sumi, Taha Yasseri, et al. Edit wars in wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 724–727. IEEE, 2011.

[180] Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on wikipedia. In *Proceedings of the Conference of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[181] Dhaval Taunk, Shivprasad Sagare, Anupam Patil, Shivansh Subramanian, Manish Gupta, and Vasudeva Varma. Xwikigen: Cross-lingual summarization for encyclopedic text generation in low resource languages. In *Proceedings of the ACM Web Conference 2023*, pages 1703–1713, 2023.

[182] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

[183] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[184] Gerrit JJ Van den Burg and Christopher KI Williams. An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*, 2020.

[185] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*, 2019.

[186] Fernanda B Viegas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. Talk before you type: Coordination in wikipedia. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 78–78. IEEE, 2007.

[187] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*, 2015.

[188] Meihong Wang, Linling Qiu, and Xiaoli Wang. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3):485, 2021.

[189] Ping Wang and Xiaodan Li. Assessing the quality of information on wikipedia: A deep-learning approach. *Journal of the Association for Information Science and Technology*, 71(1):16–28, 2020.

[190] Ping Wang, Xiaodan Li, and Renli Wu. A deep learning-based quality assessment model of collaboratively edited documents: A case study of wikipedia. *Journal of information science*, 47(2):176–191, 2021.

[191] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE transactions on knowledge and data engineering*, 29(12):2724–2743, 2017.

[192] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat-Seng Chua. Learning intents behind interactions with knowledge graph for recommendation. In *Proceedings of the Web Conference 2021*, pages 878–887, 2021.

[193] Yu Wang and Tyler Derr. Degree-related bias in link prediction. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 757–758. IEEE, 2022.

[194] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[195] Morten Warncke-Wang, Dan Cosley, and John Riedl. Tell me more: an actionable quality model for wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, pages 1–10, 2013.

[196] Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 743–756, 2015.

[197] Dennis M Wilkinson and Bernardo A Huberman. Assessing the value of coooperation in wikipedia. *arXiv preprint cs/0702140*, 2007.

[198] Thomas Wöhner and Ralf Peters. Assessing the quality of wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, pages 1–10, 2009.

[199] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 975–985, 2016.

[200] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[201] Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*, 2015.

[202] Diyi Yang, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. Who did what: Editor role identification in wikipedia. In *ICWSM*, pages 446–455, 2016.

[203] Olga Zagovora, Fabian Flöck, and Claudia Wagner. ”(weitergeleitet von journalistin)” the gendered presentation of professions on wikipedia. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 83–92, 2017.

[204] Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel P Robert Jr, and Daniel M Romero. Crowd development: The interplay between crowd evaluation and collaborative dynamics in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–21, 2017.

[205] Charles Chuankai Zhang and Loren Terveen. Quantifying the gap: a case study of wikidata gender disparities. In *Proceedings of the 17th International Symposium on Open Collaboration*, pages 1–12, 2021.

[206] Haifeng Zhang, Yuqin Ren, and Robert E Kraut. Mining and predicting temporal patterns in the quality evolution of wikipedia articles. In *HICSS*, pages 1–10, 2020.

[207] Shiyue Zhang, Zheng Hu, Chunhong Zhang, and Ke Yu. History-based article quality assessment on wikipedia. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8. IEEE, 2018.

[208] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[209] Wen Zhang, Bibek Paudel, Wei Zhang, Abraham Bernstein, and Huajun Chen. Interaction embeddings for prediction and explanation in knowledge graphs. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 96–104, 2019.

[210] Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3065–3072, 2020.

[211] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL 2019*, 2019.

[212] Lei Zheng, Christopher M Albano, Neev M Vora, Feng Mai, and Jeffrey V Nickerson. The roles bots play in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20, 2019.

[213] Yang Zhong. Wikibias: Detecting multi-span subjective biases in language. Master's thesis, The Ohio State University, 2021.

[214] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[215] Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288–295, 2000.

All Publications

The following is a list of publications during my tenure as a student at IIT Kharagpur. The publications are listed in chronological order.

- **Paramita Das**, Bhanu Prakash Reddy Guda, Debajit Chakraborty, Soumya Sarkar, and Animesh Mukherjee. *"When Expertise Gone Missing: Uncovering the Loss of Prolific Contributors in Wikipedia."* Towards Open and Trustworthy Digital Societies. ICADL (2021). pp. 291 – 307,
  DOI: https://doi.org/10.1007/978-3-030-91669-5_23
- **Paramita Das**, Bhanu Prakash Reddy Guda, Sasi Bhushan Seelaboyina, Soumya Sarkar and Animesh Mukherjee. *"Quality Change: Norm or Exception? Measurement, Analysis, and Detection of Quality Change in Wikipedia"*. Proceedings of the ACM on Human-Computer Interaction (2022), Volume 6, Issue CSCW1, Article No. 112, pp. 1 – 36,
  DOI: https://doi.org/10.1145/3512959
- **Paramita Das**, Sai Keerthana Karnam, Anirban Panda, Bhanu Prakash Reddy Guda, Soumya Sarkar and Animesh Mukherjee. *"Diversity matters: Robustness of bias measurements in Wikidata."* Proceedings of the 15th ACM Web Science Conference (2023), pp. 208 – 218,
  DOI: https://doi.org/10.1145/3578503.3583620
- **Paramita Das**, Isaac Johnson, Diego Saez-Trumper, Pablo Aragon. *"Language-Agnostic Modeling of Wikipedia Articles for Content Quality Assessment across Languages."* Proceedings of the International AAAI Conference on Web and Social Media (2024), Volume 18, pp. 1924 – 1934,
  DOI: https://doi.org/10.1609/icwsm.v18i1.31436
- **Paramita Das**, Amartya Roy, Ritabrata Chakraborty, and Animesh Mukherjee. *"On the effective transfer of knowledge from English to Hindi Wikipedia"*. Proceedings of the 31st International Conference on Computational Linguistics: Industry

Track (2025), pp. 453-465,
DOI: https://aclanthology.org/2025.coling-industry.39/

- **Paramita Das**, Sai Keerthana Karnam, Aditya Soni, and Animesh Mukherjee. *"Social Biases in Knowledge Representations of Wikidata separates Global North from Global South."* Proceedings of the 17th ACM Web Science Conference (2025),
DOI: https://doi.org/10.1145/3717867.3717882

- Sayantan Adak, Souvic Chakraborty, **Paramita Das**, Mithun Das, Abhisek Dash, Rima Hazra, Binny Mathew, Punyajoy Saha, Soumya Sarkar, and Animesh Mukherjee. *"Mining the online infosphere: A survey."* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2022), Volume 12, pp. e1453,
DOI: https://doi.org/10.1002/widm.1453

- Sayantan Adak, Pauras Mangesh Meher, **Paramita Das**, and Animesh Mukherjee. *"REVerSum: A Multi-staged Retrieval-Augmented Generation Method to Enhance Wikipedia Tail Biographies through Personal Narratives"*. In Proceedings of the 31st International Conference on Computational Linguistics: Industry Track (2025), pp. 732-750,
DOI: https://aclanthology.org/2025.coling-industry.61/