ASSESSING THE NEUTRALITY OF EDITS ON WIKIPEDIA

Christos Porios*
Harvard Kennedy School

Louis Guerin*
Harvard Kennedy School

Bruce Schneier
Harvard Kennedy School

Abstract

We built a custom machine learning classifier that assesses the neutrality of edits on Wikipedia, using the ongoing conflict in the Gaza strip as a case study. Our classifier determines if an edit increases, decreases, or does not affect the Neutral Point of View of an article, as per the Wikipedia definition. This study is a work in progress.

Keywords: Wikipedia, neutrality, NPOV, content bias, information quality

Introduction

Neutrality of Point of View (NPOV), or "representing fairly, proportionately, and, as far as possible, without editorial bias, all the significant views that have been published by reliable sources on a topic" is one of Wikipedia's three core content policies. Both hailed as a spectacular innovation and vindicated as a squelcher of discussion (Van Dijck, 2013), NPOV has been described by scholars as mostly aspirational.

We argue that assessing article neutrality remains essential in a context of rising disinformation and misinformation on Wikipedia. We propose a novel approach and build a classifier to evaluate the NPOV impact of individual edits. As a case study, we use English Wikipedia pages related to the ongoing conflict in the Gaza strip.

Literature review

Many scholars have grappled with Wikipedia's NPOV policy. They first looked at the process by which NPOV is achieved. Shi *et al.* (2019) argue that engaging polarized editors is effective in generating a neutral point of view. Others, like Mai (2016) argue that the homogeneity of Wikipedia's contributors makes the encyclopedia incompatible with a true neutrality of point of view.

Tools such as Contropedia were developed to assess the controversial nature of an article at a given time (Borra, *et al.*, 2014), and detailed assessments of Wikipedia's neutrality were also conducted, at the article or article corpus level. In 2013, Greenstein and Zhu studied 28,382 articles and concluded that the tendency was toward an improvement of the neutrality of articles on United States politics, due to the "entry of later vintages of articles with an opposite point of view from earlier articles". When she analyzed Wikipedia wars and the Israel-Palestine conflict, Sanbar (2021) had similar conclusions. In 2020, Góngora-Goloubintseff conducted a comparative study of the Spanish and English pages related to the Falkland/Malvinas War and concluded that because neutrality was a result of the consensus reached by the Wikipedia community, it was inherently a "local and relative position". The discrepancy between different language versions of the same article was also pointed out by Baigutanova *et al.* (2023), as they found that some sources deemed untrustworthy in English continued to appear on articles in other languages. Finally, Guo *et al.* (2023) designed Edit-History Vis, a tool to track and analyze Wikipedia edits, that however does not assess impacts on the article's NPOV.

We could not find a study that specifically analyzed NPOV at the individual edit level. This is the focus of our work. We evaluate the nature of specific Wikipedia edits (NPOV increasing, NPOV neutral, or NPOV decreasing), and track the evolution of the proportion of each of these categories over time. Our key hypothesis is that over time, in each article, NPOV decreasing edits tend to make up an increasingly lower share of new edits.

Methods

We build a classifier that classifies Wikipedia edits as either "Increases NPOV", "Decreases NPOV" or "Does not affect NPOV".

As training data, along with the Wikimedia Foundation's Trust & Safety team, we selected a list of 53 priority articles related to the ongoing war in the Gaza strip and added 9 random articles from the Wikipedia corpus, resulting in a set of 62 articles. We then extracted all edits from these articles, to obtain a set of 21,530 edits.

We (the two main authors) then manually labeled 300 random individual edits on these articles, made between February 16, 2005, and March 04, 2024. We set precise labeling rules, related to e.g., the use of biased language, the accuracy of the sources, and the article's talk paged discussions. Next, we extracted the following features for these edits:

- **Username features**: the username itself, whether the username is an IP (i.e. an anonymous contributor), and an 8-dimensional embedding of the username.
- **Diff text**: the diff text itself, and a 16-dimensional embedding of the diff text.

* Equal contributions

- **Revert risk model score**: the likelihood of the edit being reverted, according to the Wikipedia Language-Agnostic Revert Risk Model[1].
- **Past edits statistics**: the number of previous edits, and distribution statistics (mean, quantiles, standard deviation etc.) for the distribution of time differences between past subsequent edits.
- **Past user edit statistics**: same as for past edit statistics, but only for the subset of edits that were done by the user making the current edit.

All features for an edit are extracted from data that preceded the timestamp of the edit. All embeddings are generated by OpenAI's text-embedding-3-small, a model that allows us to pick the dimensionality of the resulting embedding. As a baseline model, we ask a large language model, OpenAI's GPT-4, to classify the diff text as NPOV increasing, NPOV decreasing, or NPOV neutral.

We then trained a neural network, whose hyper-parameters were picked automatically by Google's Vertex AI, on 250 training examples. The training data as well as the code for the classifier, training set generation, feature extraction and baseline comparison are available and open source on GitHub.[2]

Preliminary Results and Discussion

First, we compare our manual classification with the performance of a large language model, GPT-4. The LLM's overall accuracy is 64.5% with a weighted F1 score of 67%. A confusion matrix is available in Table 1, and performance metrics in Table 2.

Our neural network's precision recall AUC is 0.71 (curves presented in Figure 3), and at a confidence threshold of 0.25, we find a weighted F1 score of 69%. Our model, which relies on data from just 250 labels, slightly outperforms the GPT-4 baseline.

Next steps

The main output of this preliminary paper is the establishment of the methodology and training pipeline for a classifier. We are now working on improving our preliminary results by:

- **Labeling more data**: build a training dataset of 1,000 Wikipedia edits by having new labelers label an additional 700 edits manually.
- **Extracting more features**: extract additional features, including the page's ORES score, the editor's tenure and behavior outside the article being examined, as well as a reliability assessment of the sources added or removed.

Implications of our work

We identified two major use cases for our work. First, we are hoping that the classifier will be used to flag potentially NPOV-decreasing edits to the community and generate discussions on the corresponding talk pages. We will design the model to favor type-2 errors (false negatives) to only generate discussions when relevant.

Second, we will use the classifier to analyze the evolution of editing behavior over time, assessing our hypothesis that NPOV decreasing edits tend to make up an increasingly lower share of new edits. Conversely, we will study the proportional evolution of NPOV increasing edits.

References

Baigutanova, A., *et al.* Longitudinal Assessment of Reference Quality on Wikipedia. In P*roceedings of the ACM Web Conference 2023 (WWW '23), April 30--May 04, 2023, Austin, TX, USA*. ACM, New York, NY, USA (2023).

Borra, E., *et al*. *Contropedia: The Analysis and Visualization of Controversies in Wikipedia Articles*. New York: Association for Computing Machinery, 2014.

Góngora-Goloubintseff, J.G. The Falklands/Malvinas war taken to the Wikipedia realm: a multimodal discourse analysis of cross-lingual violations of the Neutral Point of View. *Palgrave Commun* 6, 59 (2020).

Greenstein, S. and Feng, Z. 2012. "Is Wikipedia Biased?" *American Economic Review*, 102 (3): 343-48.

Guo, Y., *et al*. "Edit-History Vis: An Interactive Visual Exploration and Analysis on Wikipedia Edit History," *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*, Seoul, Korea, Republic of, 2023, pp. 157-166.

Mai, J.-E. (2016). Wikipedians' Knowledge and Moral Duties. *Nordisk Tidsskrift for Informationsvidenskab Og Kulturformidling*, 5(1), 15–22.

Sanbar, S. (2021). Wikipedia Wars and the Israel-Palestine Conflict.

Shi, F., *et al*. The wisdom of polarized crowds. Nat Hum Behav 3, 329–336 (2019).

van Dijck, J., 'Wikipedia and the Neutrality Principle', *The Culture of Connectivity: A Critical History of Social Media* (Oxford Academic, 24 Jan. 2013).

Wikipedia. "Neutral Point of View". Available at: https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view (Accessed April 14, 2024)

[1] https://meta.wikimedia.org/wiki/Machine_learning_models/Proposed/Language-agnostic_revert_risk

[2] https://github.com/christosporios/wikipedia-npov-classifier/tree/main

Tables and figures

		GPT-4 predictions		
		NPOV neutral	NPOV increasing	NPOV decreasing
Manual classification	NPOV neutral	**175**	39	20
	NPOV increasing	21	**10**	11
	NPOV decreasing	10	6	**9**

Table 1: Baseline: GPT-4 predictions vs manual classification (n=300)

	Precision	Recall	F1-Score
NPOV neutral	84.9%	74.8%	79.5%
NPOV increasing	18.2%	23.8%	20.6%
NPOV decreasing	22.5%	36.0%	27.7%

Table 2: Summary of GPT-4 prediction performance (n=300)

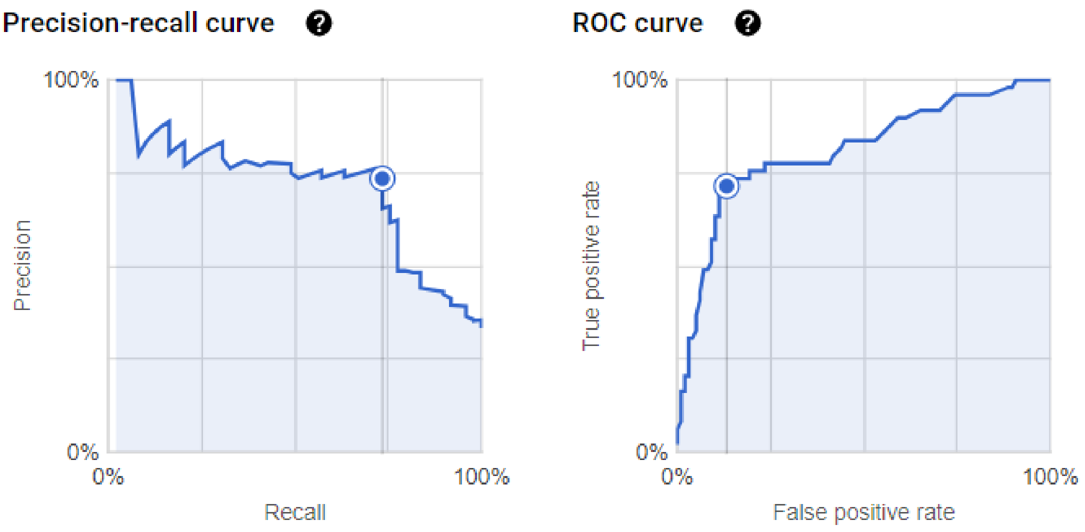

Figure 3: Evaluation metrics of a neural network trained on Vertex AI (test n=50)