Political Neutrality on Wikipedia: An Analysis of the 2020 U.S. Elections

Mayank Thakur
Delft University of Technology
Delft, The Netherlands
mthakur@tudelft.nl

Meenakshi Girish Nair
Delft University of Technology
Delft, The Netherlands
mgirishnair@tudelft.nl

Vaishnav Srinidhi
Delft University of Technology
Delft, The Netherlands
vsrinidhi@tudelft.nl

Abstract

Wikipedia is a widely used open source hypertext publication, collaboratively edited and managed by its users. One of the main policies of Wikipedia is their "neutral point of view" (NPOV), which states that all the content on Wikipedia must be written fairly and without editorial bias. It is important to maintain neutrality, as the integrity of online information is crucial in shaping people's opinion, especially during politically sensitive times like a national election. However, because Wikipedia can be edited by anyone, it is often difficult to maintain this neutrality. This paper analyzes the neutrality of relevant Wikipedia pages during the course of the 2020 U.S. Presidential Elections. In this work, distinct clusters of Wikipedia articles related to the 2020 elections are constructed using implicit and explicit relationships, and major events are identified by detecting bursts in editing activity based on the Entity Clustering Algorithm proposed by Tran et al. We observe the political stance of the edits and analyze the neutrality of these stances. The political stance is then detected using a fine-tuned BERT model and analyzed over time, assessing the reliability of the relevant articles during times of political unrest. Our use case, that is, analysis of the neutrality of Wikipedia during the 2020 U.S. presidential election, helps identify the extent to which the platform maintains neutrality in a highly fluid political situation and thus is a reflection on the reliability of the platform as an information source during crucial events.

CCS Concepts

• **Information systems → Social networks**; • **Human-centered computing → Collaborative and social computing systems and tools**.

1 Introduction

Wikipedia is one of the world's largest open source information repositories, with its content collaboratively generated and curated by a global community of volunteer editors. A foundational pillar of its editorial policy is the maintenance of a "neutral point of view" (NPOV), which mandates that articles be presented fairly, proportionately and without editorial bias [2, 18]. The adherence to NPOV is paramount, as the integrity of information on such a widely accessed platform plays a crucial role in shaping public understanding and discourse, particularly during politically sensitive events such as national elections. However, the very openness that defines Wikipedia presents a significant challenge to the consistent maintenance of this neutrality, as the platform can be susceptible to biased contributions and "edit wars", especially on contentious topics [12, 20].

While the strength of Wikipedia lies in its capacity for rapid, collaborative updates, this dynamic environment creates an inherent tension with its NPOV policy. Previous research has highlighted several systemic challenges to neutrality, including the disproportionate influence of a small group of "elite users" [6, 10], the emergence of bias among administrators [1], and the prevalence of cultural and gender biases across the platform [14]. These challenges are exacerbated during periods of intense political activity, such as the 2020 U.S. Presidential Election. The rapid succession of events, coupled with profound political polarization, creates a fertile ground for biased editing, potentially diminishing the reliability of Wikipedia articles when a neutral source of information is needed most [15]. While the existence of bias on Wikipedia is established, there remains a need for a granular analysis of how neutrality fluctuates in real-time in response to specific, high-impact events as they unfold.

This paper presents an empirical analysis of Wikipedia's political neutrality throughout the course of the 2020 U.S. Presidential Election. We employ a two-stage methodology to investigate the relationship between real-world events and the political stance of content on the platform. First, we systematically identify significant election-related events by detecting "bursts" in editing activity across a custom-built dataset of relevant Wikipedia articles, adapting a method for event detection from Wikipedia's edit history [13]. Second, following the identification of these events, we analyze the content of the corresponding edits to assess their political leaning. For this, we utilize a BERT-based language model, fine-tuned on the P-Stance dataset [7], to classify the political stance of individual edits as favoring or opposing the primary presidential candidates.

Through this investigation, we aim to answer the following research questions:

(1) How can significant real-world events related to the 2020 US election be systematically identified by analyzing bursts of editing activity on relevant Wikipedia pages?
(2) What is the political stance of the edits that occur during these event-driven bursts, and how does an analysis of this stance inform our understanding of Wikipedia's adherence to its NPOV policy during a period of intense political conflict?

By examining the interplay between offline events and online content dynamics, this study contributes a quantitative assessment of Wikipedia's neutrality during a critical political moment. The

DSAIT4055-Q4-25, TU Delft

findings offer insights into the resilience of Wikipedia's collaborative moderation model and its reliability as a source of neutral information in a highly polarized environment.

2 Related Work

Our research builds upon several key areas of prior work: the challenges to Wikipedia's Neutral Point of View (NPOV) policy, methods for detecting real-world events from online data, and the application of Natural Language Processing (NLP) for stance detection.

2.1 Challenges to Wikipedia's NPOV Policy

Wikipedia's NPOV policy aims to represent all significant views proportionately, attribute opinions to their sources, and employ nonjudgmental language [18]. Despite being a "non-negotiable principle", the practical application of NPOV is fraught with challenges, especially during politically sensitive times such as national elections [15], where the reliability of Wikipedia articles can diminish. The policy itself acknowledges that articles should be written "as far as possible" without editorial bias [17]. Academic studies confirm that Wikipedia articles, while often factually accurate, can be incomplete and may not present all perspectives with a truly neutral point of view. This highlights a discernible gap between the policy's intent and its real-world execution, especially when covering "fast-moving and contentious topics" [16].

The main challenges, identified by prior research, to maintaining neutrality are:

- Influence of Elite Users: [10, 16] found that merely 0.1% of the Wikipedia community generated 44% of the persistent word views over a four-year period. Similarly, [6] observed "social stratification" within the Wikipedia community, noting that while the direct influence of administrator edits diminished, "elite users" (those with over 10,000 edits) continued to contribute substantially to content development. This observed social stratification can implicitly shape the evolution and perspective of content.
- Administrator Bias: [1] developed a statistical model to measure systematic bias in user behavior, finding that bias emerged after administrators assumed their posts, suggesting a potential for manipulation among these "arbiters of collective intelligence".
- Systematic Bias: Beyond the adherence to NPOV, academic research has identified various forms of systemic bias within Wikipedia. These include cultural and gender biases, as well as notable gaps in the coverage of the Global South. For instance [14]
- Edit Wars: Wikipedia articles on controversial topics are prone to "edit wars". [12] modeled social dynamics in collaborative environments, observing that "a peaceful article can suddenly become controversial when more people get involved in its editing".
- Times of conflict: [20] directly investigated Wikipedia's neutrality from a sentiment perspective, focusing on war-related articles since 1945. Their study provides a crucial methodological precedent for assessing neutrality through linguistic analysis, which is highly pertinent to the current project's proposed use of sentiment and stance detection.

The strength of Wikipedia lies in its open and collaborative editing environment, which facilitates diverse contributions and rapid information updates. Yet, this openness poses a fundamental challenge to adherence to the NPOV policy. This body of work establishes the foundational problem that our study addresses: the inherent tension between Wikipedia's NPOV policy and the realities of its open, collaborative environment. Our research extends this prior work by moving beyond identifying the existence of bias to quantitatively analyzing how neutrality fluctuates in response to specific, high-impact political events, namely the 2020 U.S. Presidential Election.

2.2 Event Detection Using Wikipedia's Edit History

The identification of significant events from continuous streams of online data is a critical area of research, particularly as user-generated content platforms proliferate. These "remarkable happenings" can range from large-scale societal occurrences, such as natural disasters or national elections, to more localized incidents like traffic congestion.

Event detection methodologies typically fall into supervised, unsupervised, or hybrid categories. Unsupervised methods, such as online clustering, are particularly valuable for "new event detection" where the specific event to be identified is not predetermined. These methods group continuously created text streams into topic clusters based on their similarity over time [5]. While general social media data, such as tweets, are frequently employed for event detection, Wikipedia's edit history offers a distinct advantage. Unlike the often fleeting nature of social media posts, Wikipedia edits represent a collaborative, frequently deliberative, and persistent record of information evolution. Bursts in editing activity on Wikipedia pages often signify a collective response to real-world events, as editors actively update and refine information. This characteristic makes Wikipedia's edit history a more structured and potentially more reliable signal for identifying significant public events compared to the often noisy nature of general social media.

The current work specifically aims to detect events during the 2020 U.S. Elections by identifying bursts in editing activity on Wikipedia pages. This approach is directly inspired by existing research that leverages Wikipedia's comprehensive edit history. A primary methodological inspiration for the project's event detection component is from the work by Tran et al. [13] that provides foundational techniques for using the granular history of Wikipedia edits to identify significant occurrences, likely involving analysis of patterns and anomalies in edit frequency and content changes. Additionally, Meier's work also inspires the automatic classification of events using edit history [8]. This suggests that techniques for identifying "bursts" or significant changes in user engagement, whether through views or edits, can be generalized to detect events.

Temporal analysis is an indispensable component of event detection, as events are inherently characterized by their temporal dimension. Real-time event detection specifically focuses on identifying happenings as they unfold, necessitating continuous monitoring and analysis of incoming data streams. Various methodologies leverage temporal information for different types of event detection. For instance, in traffic accident detection, explicit temporal information from social media posts is utilized. For new event detection,

incremental clustering techniques are well-suited for continuously created text, enabling the categorization of text streams into topic clusters based on their similarity over time [5]. Deep learning approaches, such as BiLSTM, have been employed to extract time- and location-related information for emergency events [9]. The project's methodology explicitly links event detection, identified as "times of conflict," to a subsequent neutrality assessment. This establishes a clear relationship: significant real-world events, detected through bursts in editing activity, are hypothesized to cause shifts or challenges in Wikipedia's neutrality. This represents a crucial analytical step, extending beyond mere event identification to understanding the impact of these events on information quality. The temporal analysis of neutrality, specifically examining the state of pages before and after the edits, directly supports this investigation into cause and effect.

2.3 Stance and Bias Detection using NLP

Sentiment analysis, also known as opinion mining, is a well established NLP task focused on distinguishing subjective information in text, such as positive or negative sentiment. Early research in media bias detection often relied on hand-developed lexical and linguistic techniques, alongside sentiment analysis and topic modeling. While these lexical-based models proved effective for identifying explicit word choice bias, they often fell short in detecting more complex forms of bias, such as persuasion or rhetoric [11]. This limitation has been a primary driver for the shift towards more advanced neural methods in the field. Stance detection represents a more nuanced NLP task compared to general sentiment analysis. Its objective is to categorize the stance—whether agreement, disagreement, support, opposition, or neutrality—taken by a text toward specified targets or claims [4]. This capability is widely applied in critical areas such as misinformation detection, political opinion mining, and the verification of fake news. Political discourse, in particular, offers a rich ground for stance analysis, as comments within this domain are frequently "laden with sentiment, which can serve as an implicit indicator of stance" [19]. Jin and Mihalcea highlight that NLP methods are increasingly being applied to political science for tasks including identifying political perspectives in news articles, analyzing campaign sentiment, and categorizing legislative issues [4].

The emergence of deep learning, particularly transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), has revolutionized the field of NLP. These models possess the ability to automatically learn rich feature representations directly from text, consistently outperforming traditional methods in various NLP tasks, including sentiment and stance detection. The current project specifically proposes utilizing a fine-tuned BERT model to classify support and/or rejection of individuals relevant to the elections and to assess the neutrality of Wikipedia edits and pages. This methodological choice aligns with the broader trend of leveraging powerful pre-trained language models for complex text analysis.

While BERT is a highly capable tool for text classification, including stance detection, it is important to acknowledge its position within the continuously evolving landscape of NLP models. For instance, the PoliStance-VAE framework has explicitly demonstrated state-of-the-art performance, surpassing models like BERT, BERTweet, and GPT-4o in political stance detection [19]. The success of PoliStance-VAE stems from its ability to disentangle latent emotional features (Valence, Arousal, and Dominance - VAD) and integrate nuanced sentiment information, aspects that standard BERT fine-tuning might not fully capture. This indicates that while BERT remains a strong and valid baseline, more advanced models are continually emerging that specifically address the intricate complexities of political stance. The availability of high-quality, domain-specific datasets is paramount for effectively fine-tuning neural models for tasks such as stance detection. The P-Stance dataset by Li et al. is explicitly identified as a key enabler for the current project, while also being a limiting factor for its applicability to other time periods [7]. This dataset provides the necessary labeled data for training the BERT model to classify political support or rejection, making it fundamental to the project's neutrality assessment methodology. The existence of such a specialized dataset allows the project to move beyond general sentiment analysis to a more specific and granular political stance classification.

Detecting bias in online information sources presents a significant challenge due to the inherent complexity and ambiguity of human language [11]. Recent advancements in this area increasingly utilize neural network models, including Recurrent Neural Networks (RNNs) and transformers, which possess the capability to automatically learn sophisticated feature representations from text. Beyond merely identifying lexical bias, more advanced models, such as the cascade transformer system discussed by Rodrigo-Ginés et al., aim to detect specific forms of bias, including persuasive techniques like appeals to emotion, authority, or loaded language. This distinction is crucial because neutrality can be determined not only by explicit biased words but also by subtle rhetorical strategies embedded within the text.

A critical consideration when employing models like BERT is their inherent susceptibility to biases [3]. Large Language Models (LLMs), including BERT, are trained on vast datasets, often sourced from the internet, and inevitably inherit biases present in these foundational sources. These biases can be categorized as intrinsic, stemming from the training data or model architecture, or extrinsic, manifesting in the model's performance on specific tasks. This creates a critical scenario where the analytical instrument's own predispositions could potentially skew the assessment of Wikipedia's neutrality. Comprehensive reviews on LLM bias discuss origins, evaluation methods, and mitigation strategies, emphasizing the need for robust frameworks to detect, quantify, and mitigate these biases to ensure LLMs function as fair and equitable tools [3].

3 Method

3.1 Dataset

n order to investigate the neutrality of Wikipedia, we use two datasets. One, the custom dataset that we create for event detection during the course of 2020 U.S. presidential elections and another one to finetune the BERT model for stance detection to answer the second research question.

3.2 Custom Dataset

The custom dataset was created by scraping Wikipedia. The dataset consists of Wikipedia pages, their edit histories over certain time periods, and all the pages they link to. One of the biggest challenges

of scraping from a database like Wikipedia is relevance. Since we did not have the resources to exhaustively collect all of the relevant pages, we implemented a pruned breadth first search (BFS), starting from a set of nucleation pages. The nucleation pages are an initial set of Wikipedia pages that we know are relevant to the 2020 U.S. Election, from which the scraper will find the rest of the pages. The selected pages are listed in Appendix A. The searchable space of BFS grows exponentially, as with just a depth of 2, we had to analyze over 50,000 articles. In order to increase the usefulness of the dataset, we only stored pages who's first paragraphs had a BERT embeddings had a cosine similarity of more than 0.8.

The next challenge was to gather edit histories for all the relevant articles. We used the WikiMedia API to get the edit histories and organized them according to the timestamp and content of the edit for each article. Collecting edits for an entire year proved to be challenging due to time constraints and the rate limiting applied to the API calls. Therefore, we narrowed our scope to three major real-world events within January 2020 and February 2021. we sought to execute our proposed methodology within these time frames. These are:

(1) The Presidential Debate: September 1 - November 1, 2020
(2) The Election: November 3 - December 3, 2020
(3) The January 6th riots: January 1 - March 1, 2021

Since this is a retrospective study, the exact dates of all the studied events are available, but data is collected using a buffer on the known dates. This is because there is a lag between events happening and them being updated on Wikipedia. In the case of the Presidential Debate and the Election, it makes sense to have more days preceding and succeeding the event, because smaller events can occur in relation to, or in anticipation of the big event. However, in the case of the January 6th Riots, there is very minimal anticipation of it, so data preceding the event is not very useful.

Based on the context of the study, there are certain events which are not useful. For big politician pages, there are many "prank" edits, where contents of the page (sometimes the entire page) is replaced with unrelated content. These edits are usually reverted immediately. Therefore, both the edit and its reversal are not useful for event detection. Additionally, cosmetic edits, which add/ change images are also not useful for event detection. These edits will not be saved. The stance detection model takes as input only textual data. Hence, the edits are filtered to only include textual data and disregard any tables or comments.

A more "unsupervised" approach to this study would be the one provided in [13], where the authors gather the full edit history of the Current Events Wikipedia page. This page is updated everyday and can serve as a ground truth for the happenings of everyday. However, since we choose predefined time periods to conduct our analysis, there is no need to have a ground truth dataset. The study presented in this paper functions more like a toy-experiment for the method.

3.3 Event Detection

The aim of the custom dataset was to enable running the Entity Cluster Aggregation (ECA) Algorithm. Here, the entities are the Wikipedia pages. The algorithm runs at every timestep on an implicit and explicit graph. These are constructed as follows,

- **Explicit Graph:** Entities are nodes, and edges are created when there is a hyperlink from one page to another within their edit history.
- **Implicit Graph:** Constructed by calculating the cosine similarity between the BERT embeddings of the first paragraphs of the articles. Pages with similarity greater than 0.8 are connected.

Once the graphs are built, the algorithm tracks the edit frequency for each cluster over time. Bursts in editing activity within clusters are interpreted as possible real-world events. The intuition is that related articles experience a correlated increase in editing frequency when notable events occur. This co-burst behavior is a strong signal for identifying significant real-world happenings. These clusters are then filtered using thresholds on burst size and graph connectivity to identify the most plausible events.

We applied the ECA algorithm during three key periods: the Presidential Debates, the Election, and the January 6th Riots. The output was a set of detected clusters for each time window, each of which represents a candidate event.

3.4 P-Stance

P-Stance is a large stance detection dataset [7]. It is made specifically for the political domain and contains 21,574 labeled tweets. These are tweets labeled *Favor* or *Against* one of the three 2020 U.S. Presidential election candidates: Donald Trump, Joe Biden and Bernie Sanders. These tweets are collected between February and October 2020. These tweets were captured via the Twitter API by filtering the tweets for "favor" or "against". From the 2 million tweets that remained after language filtering and length constraints, 30,000 were annotated by three U.S. crowd workers. The benchmark results of this dataset includes a macro-F1 of 80.53% on the two-class test sets. Since our research is aimed at finding the political stance of Wikipedia edits during the 2020 U.S. election, fine-tuning the BERT model on this dataset should improve its performance on the Wikipedia edits. The dataset is also structured in a way that there are no hashtags, the data is just plain text, which is also beneficial for our research.

3.5 Models

In order to detect the political stance of the Wikipedia edits, a finetuned BERT model will be used. The model is finetuned on the P-Stance dataset. Since BERT is already pre-trained on a very large general domain corpus that includes Wikipedia, it already encodes the stylistic and topical knowledge of the language used in Wikipedia. This reduces domain mismatch and also allows effective fine-tuning with a small number of manually labeled edits. Additionally, the same architecture has shown state of the art performance on the P-Stance benchmark, outperforming CNN and BiLSTM.

4 Procedure

4.1 RQ1

In order to answer RQ1, the following steps were taken:

(1) Create the custom dataset.
(2) Apply the Entity Cluster Aggregation Algorithm from [13] for event detection (but with pre-defined time range around the three events we want to focus on).

4.2 RQ2

In order to answer RQ2, the following steps were taken:

(1) Get the edit histories of the Wikipedia pages per event detected.
(2) Carry out stance detection of the edits towards the three U.S. presidential candidates using the finetuned BERT model.
(3) Analyze the political stance of the edits and draw conclusions.

5 Results

5.1 Event Detection

Figure 4 depicts the event clustering algorithm. It shows candidate clusters that are formed using co-burst detection, which are then filtered to find clusters of pages that form actual events, using the methods of [13]. There are many events which are found, but the validity of the detected events is questionable. In [13], the authors cited an accuracy of 70%. However, the events detected do not see to correlate. For example, one of the events detected for the January 6 riots is trumps inauguration of the US ambassador to Tokyo, and some edits on the biography of a US politician. These may be related by BeRT embeddings, yes, but they don't add to our understanding of the event. Furthermore, there are not many events detected per time period, there are up to 15 events detected for the larger time periods. Combined with a 70% peak (possible) accuracy, it is possible that none of the events are correct events.

5.2 Stance Detection

Once the events are detected, we take all the edit per event and run it through the BERT model to obtain the stance of the edit for the three candidates: Donald Trump, Joe Biden and Bernie Sanders. Figure 1 shows the stance for each candidate over the course of September 2021 to November 2021. This time frame corresponds to when the 2020 Presidential Debate took place. The x-axis depicts the timestamp of each event. The y-axis shows the two stance predicted by the model, which is 1 (Favor) and 0 (Against). The graph is color coded for the different candidates, showing the stance against them for each edit. However, this is not visible in the graph due to the stance predicted by the model for each candidate being the exact same.

Similarly Figure 2 and Figure 3 shows the stance across edits for

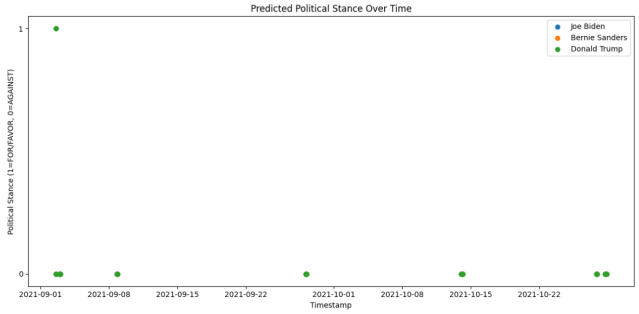

Figure 1: Political Stance over time for the event: 2020 Presidential Debate

the event January 6th riots and The Election respectively. The same pattern can be seen for both these graphs as well, with the stance for all the candidates being the same.

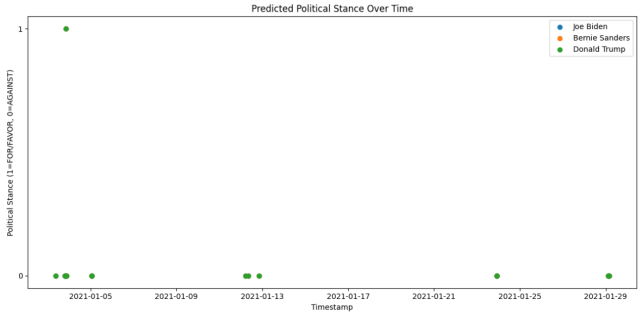

Figure 2: Political Stance over time for the event: January 6th Riots

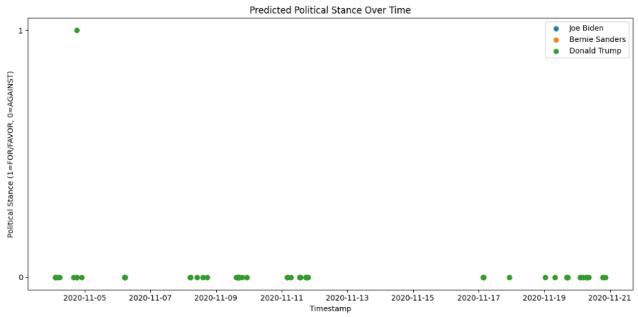

Figure 3: Political Stance over time for the event: 2020 Presidential Election

6 Analysis

6.1 Event Detection

While events are being detected, the quality of the events detected are rather low. This combined with the fact that the predicted stances are not very useful, points to the fact that the issue with event detection lies in the data collection process. Compared to [13] and other related works, the data collection process was heavily modified to reduce the computational load of the project. However, it is possible that too much data was pruned before it was analyzed. The method presented in [13] has a very general mathematical formulation, which does not seem to require a data pruning step like we had. Furthermore, there is a critical philosophical oversight within our methodology, which has to do with the unsupervised nature of the method. The goal of the project is to do unsupervised event detection and use that to gauge political stances over time. In the attempt to reduce computational load, we narrowed the dataset time periods to periods we know *should* have events. However this does not account for the lag in information between real life and Wikipedia. Putting such a constraint on the algorithm might also have limited the data collection process in other ways which also negatively impacted the results. It seems that the lack of quality in

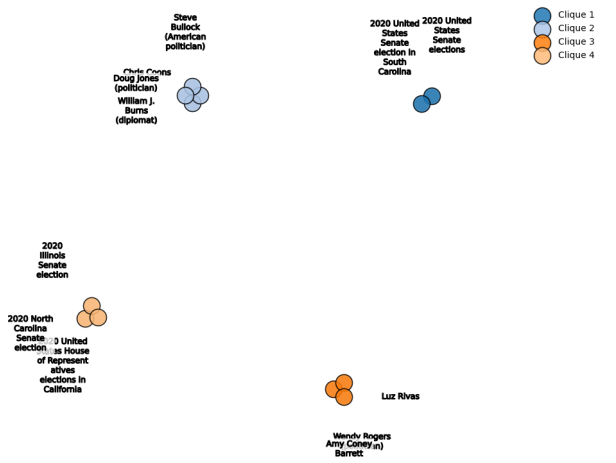

Figure 4: Implicit cliques found using the event detection algorithm

the detected events stems from the methodological choices made in the data collection phase. This is something that can be considered for a future work.

6.2 Stance Detection

The results obtained by running the BeRT model for stance detection does not show any variance. This could be the result of multiple factors. After conducting critical analysis of the results, it seems more likely that the issue is with the data we created. The BERT model is trained on the P-Stance dataset and it achieves an accuracy of around 80% on the test set. Hence, there are less chances that the model is not learning well. This points to the fact that the issue lies with the data we give the model.

Further analysis of the data leads to potential limitations of our data collection process. As mentioned in 5.1, the events detected do not have a high correlation. This would mean that some edits in an event would not have any link to the target candidates. This would make it difficult for BeRT to detect the stance of the edit or it would detect the stance to be neutral, not favoring or against the target candidates. Since the BeRT model only predicts two labels: Favor and Against, the neutral stance might be misclassified as Against. Hence, when this model receives an edit that is neutral, it will be forced into one of the two labels. This could be determined by the majority class in the training data, which for P-Stance is the "Against" class. This is seen in the results obtained as a large majority of the predictions made by BeRT on the edits are "Against" the target.

Another limitation that could have affected the predictions of BeRT could be the format of the data we extract from Wikipedia. BeRT trained on P-Stance expects plain text data as input. Although the edits were filtered to only have textual data and not tables, or images or other non-textual edits,the collected edits included raw Wikipedia *diff* markups. These tokens could have split into unknown sub-words that drowned the limited stance cues that existed in the edits.

7 Conclusions

This project attempted to quantify Wikipedia's political neutrality in the context of the 2020 U.S. Presidential Election. We designed a two-stage pipeline: (1) event detection based on Wikipedia edit bursts using a graph clustering approach, and (2) political stance classification using a fine-tuned BERT model on the P-Stance dataset. While the methodology was theoretically sound and drew from strong prior work, the results did not show strong variance in political stance across the edits.

This outcome is likely due to several constraints in the dataset construction and data filtering steps, which may have significantly reduced the quality and diversity of the collected edits. Furthermore, issues in the edit formatting and the forced binary stance classification likely introduced biases or errors in prediction. Despite these challenges, the project sheds light on the practical issues in applying NLP to real-world collaborative platforms and opens doors for improved methodologies in future work.

7.1 Broader Impact

Wikipedia is one of the most visited information sources on the internet, especially during politically charged periods like national elections. Ensuring and verifying its neutrality is crucial to fostering an informed electorate. This project contributes toward that goal by attempting to detect and evaluate political stance through machine learning and data mining techniques.

The research also serves as a cautionary tale about the practical limitations of real-world NLP applications. Although tools like BERT and stance detection models are powerful, their outputs are only as reliable as the input data and the framing of the task. Future research in this space must balance computational efficiency with methodological rigor, especially when the social implications are as significant as neutrality in public information platforms.

8 Improvements and Future works

There are several areas where this project can be improved:

- Data Collection: The pruned BFS approach limited the dataset to a small set of articles. Expanding the crawl and including multiple hops with dynamic filtering could improve coverage.
- Edit Processing: Removing Wikipedia diff markups and standardizing the input format can help reduce noise and improve the stance detection results.
- Multi-class Classification: Including a "Neutral" class in stance classification would likely yield more meaningful distinctions, especially given the nature of encyclopedic content.
- Better Ground Truth: Using external datasets such as news coverage or annotated Wikipedia events could help verify the accuracy of detected bursts.
- Alternative Models: Exploring recent models such as PoliStance-VAE that capture affective dimensions of language could provide richer stance understanding.

In conclusion, while the initial attempt provided limited insights, it lays the groundwork for a more scalable and accurate system for evaluating neutrality on Wikipedia during real-world events.

9 Link to Repository

The source code for this project is available at https://github.com/V41SH/WikiStance.

References

[1] Sanmay Das, Allen Lavoie, and Malik Magdon-Ismail. 2016. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *ACM Transactions on the Web (TWEB)* 10, 4 (2016), 1–25.

[2] Shane Greenstein and Feng Zhu. 2012. *Collective intelligence and neutral point of view: the case of Wikipedia.* Technical Report. National Bureau of Economic Research.

[3] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915* (2024).

[4] Zhijing Jin and Rada Mihalcea. 2023. Natural language processing for policymaking. In *Handbook of computational social science for policy*. Springer, 141–162.

[5] Mohammadsepehr Karimiziarani. 2022. A tutorial on event detection using social media data analysis: Applications, challenges, and open problems. *arXiv preprint arXiv:2207.03997* (2022).

[6] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web* 1, 2 (2007), 19.

[7] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-Stance: A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2355–2365.

[8] Florian Meier. 2024. Using Wikipedia Pageview Data to Investigate Public Interest in Climate Change at a Global Scale. In *Proceedings of the ACM Web Science Conference (WebSci '24)*. ACM, 1–11. doi:10.1145/3614419.3644007

[9] Motahara Sabah Mredula, Noyon Dey, Md Sazzadur Rahman, Imtiaz Mahmud, and You-Ze Cho. 2022. A review on the trends in event detection by analyzing social media platforms' data. *Sensors* 22, 12 (2022), 4531.

[10] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '07)*. Association for Computing Machinery, New York, NY, USA, 259–268. doi:10.1145/1316624.1316663

[11] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2023. Identifying media bias beyond words: using automatic identification of persuasive techniques for media bias detection. *Procesamiento del Lenguaje Natural* 71 (2023), 179–190.

[12] János Török, Gerardo Iñiguez, Taha Yasseri, Maxi San Miguel, Kimmo Kaski, and János Kertész. 2013. Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical review letters* 110, 8 (2013), 088701.

[13] Tuan Tran, Andrea Ceroni, Mihai Georgescu, Kaweh Djafari Naini, and Marco Fisichella. 2014. Wikipevent: Leveraging wikipedia edit history for event detection. In *International Conference on Web Information Systems Engineering*. Springer, 90–108.

[14] Francesca Tripodi. 2023. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New media & society* 25, 7 (2023), 1687–1707.

[15] Wikimedia Foundation. [n. d.]. Strengthening Wikipedia's neutral point of view. https://diff.wikimedia.org/2025/03/27/strengthening-wikipedias-neutral-point-of-view/. Online; accessed 11-June-2025.

[16] Wikipedia contributors. 2025. Academic studies about Wikipedia — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Academic_studies_about_Wikipedia&oldid=1294369000. [Online; accessed 11-June-2025].

[17] Wikipedia contributors. 2025. Criticism of Wikipedia — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Criticism_of_Wikipedia&oldid=1294572729. [Online; accessed 11-June-2025].

[18] Wikipedia contributors. 2025. Neutral point of view — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Neutral_point_of_view&oldid=1276201616. [Online; accessed 11-June-2025].

[19] Beiyu Xu, Zhiwei Liu, and Sophia Ananiadou. 2025. Disentangled VAD Representations via a Variational Framework for Political Stance Detection. *arXiv preprint arXiv:2502.19276* (2025).

[20] Yiwei Zhou, Alexandra I. Cristea, and Zachary Roberts. 2015. Is Wikipedia Really Neutral? A Sentiment Perspective Study of War-related Wikipedia Articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)*. 160–168.

A List of Nucleation Pages

- 2020 United States presidential election
- Donald Trump
- Joe Biden
- Kamala Harris
- Mike Pence
- 2020 Republican Party presidential primaries
- 2020 Democratic Party presidential primaries
- United States presidential debates, 2020
- 2020 Republican National Convention
- 2020 Democratic National Convention
- Donald Trump 2020 presidential campaign
- Joe Biden 2020 presidential campaign
- Opinion polling for the 2020 United States presidential election
- Russian interference in the 2020 United States elections
- Impact of the COVID-19 pandemic on the 2020 United States presidential election
- Postal voting in the 2020 United States elections
- Endorsements in the 2020 United States presidential election
- Timeline of the 2020 United States presidential election
- Political positions of Joe Biden
- Political positions of Donald Trump

Received 20 June 2025