Sentiment Analysis on Wikipedia RfA Network Data

Taeesh Azal Assadi
Virginia Polytechnic Institute and State University
Falls Church, Virginia, United States
taeeshazalassadi@vt.edu

Faraz Ulhaq Shah
Virginia Polytechnic Institute and State University
Falls Church, Virginia, United States
shahfaraz@vt.edu

ABSTRACT

The Wikipedia link graphs (SNAP) dataset is a significant collection of Wikipedia articles and their linkages. The dataset has been used for a variety of tasks, including information retrieval, natural language processing, and social network analysis. In this study, we analyze a portion of the Wikipedia RfA network data, Sentiment analysis on the data was done using sophisticated natural language processing techniques. The results showed that the RfA comments were categorized as positive, negative, or neutral, offering information about the network dynamics. In addition, These observations provide insightful data on the components of successful adminship requests inside the RfA network.

KEYWORDS

Wikipedia, sentiment analysis, social network analysis, natural language processing, RfA, community governance

ACM Reference Format:
Taeesh Azal Assadi and Faraz Ulhaq Shah. 2024. Sentiment Analysis on Wikipedia RfA Network Data. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Wikipedia is a free online encyclopedia that has contributors from all around the world. With more than 6 billion monthly users, it is one of the most well-known websites in the entire globe. Wikipedia is a great source of knowledge on a variety of subjects. It is, however, a system that is complicated and dynamic. Making sure the content on Wikipedia is correct and current is one of the maintenance tasks. Making the community welcoming and inclusive is a further difficulty. Wikipedia has several policies and procedures in place to handle these issues. The Request for Adminship (RfA) procedure is one of these policies. The RfA procedure is a method by which editors can ask to be given administrative authority on Wikipedia. These duties include managing disputes, deleting content, and blocking users. To be successful in a RfA request, an editor must show they have the abilities, expertise, and understanding required to complete these duties.

The RfA procedure plays a significant role in preserving Wikipedia's caliber and reliability. It is, however, a difficult and drawn-out process. In this study, we examine some of the Wikipedia RfA network data using techniques for natural language processing. To find trends, we apply sentiment analysis and, It's possible to examine the dynamics of this process using the RfA network data from the Wikipedia link graphs dataset. The network data contains remarks made by participants in the RfA process, providing information about the mood, themes, and dynamics of the network. The examination of this information can aid in understanding editor behavior, background, and contributions as well as the elements that contribute to successful admin requests.

Knowing the RfA procedure is crucial for the success of the Wikipedia community because admins are responsible for upholding the accuracy and reliability of the content. Indicators of the health of the community can be found in the administration process since it reflects the values, expectations, and standards of the community. It is crucial to identify the essential factors that result in successful R fA requests to improve the effectiveness and efficiency of the Wikipedia RfA (Request for Adminship) process. Understanding the components that lead to success is essential because the process is complex and time-consuming. The RfA procedure can be made more efficient by identifying these characteristics, and candidates will be better equipped to meet their requirements. The RfA process on Wikipedia is essential for maintaining the platform's reputation and quality. The process must be improved in terms of efficacy and efficiency by identifying the critical elements that lead to successful RfA requests. This will guarantee that Wikipedia remains a trustworthy source of knowledge on a variety of subjects.

The Wikipedia RfA process is important to several stakeholders, including:

- **Wikipedia editors:** The elements that go into making a successful RfA request must be understood by editors who want to become administrators.
- **Users of Wikipedia:** Users of Wikipedia need to be satisfied that the site's administrators are competent to do so.
- **The Wikimedia Foundation:** The non-profit organization that runs Wikipedia is known as the Wikimedia Foundation. The Foundation wants to increase the RfA process' efficacy and efficiency.

To gain insights into the factors that contribute to successful RfA (Request for Adminship) requests on Wikipedia, we will analyze a portion of the RfA network data using natural language processing techniques. Specifically, we will utilize sentiment analysis to identify patterns within the RfA comments. By doing so, we aim to derive meaningful findings that shed light on the dynamics of successful RfA requests. Our overall plan for approaching the problem is as follows:

- Collect a dataset of Wikipedia RfA.
- Use natural language processing techniques to analyze the dataset.
- Identify patterns in the RfA comments.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- Analyze the patterns to identify the factors that contribute to successful RfA requests.
- Communicate the findings to the relevant stakeholders.

We think that this study will significantly advance knowledge of the Wikipedia RfA procedure. All stakeholders will gain from our results, which will contribute to the process's increased effectiveness and efficiency.

2 RELATED WORKS

In the paper by Torunoglu et al. [1], the authors propose a method for enhancing sentiment classification accuracy in Twitter data using Wikipedia sentiment knowledge. They utilize semantic smoothing to find relevant words for sentiment classification. Twitter data is preprocessed before being used for sentiment analysis. The proposed method shows potential to improve sentiment classification performance not only on Twitter but also on other social media platforms. The feature set generated in this method has been noted to perform better than the baseline Twitter dataset, yielding improved accuracy.

Yang et al. [2] analyze whether Wikipedia articles on controversial topics like war maintain neutrality by studying their sentimental orientation over time. They collected a dataset of 243 Wikipedia articles related to war and applied sentiment analysis. The study found a correlation between the number of people involved in editing and the empathy score for the content. Additionally, the authors proposed a framework for analyzing sentiments in bilingual articles.

Bhuta et al. [3] provide an overview of various techniques for sentiment analysis of Twitter data. In their paper, the authors explore and experiment with different machine learning and deep learning methods for sentiment analysis, highlighting their advantages and disadvantages. They also investigate the benefits of using different datasets for sentiment analysis tasks.

In the paper by Tiwari et al. [4], the authors focus on identifying the emotional tone of tweets through sentiment analysis using machine learning techniques. They utilized the Indian Premier League dataset for their analysis and compared the performance of multiple machine learning algorithms. The study concluded that selecting the right feature set and algorithm significantly impacts the results, improving the effectiveness of sentiment analysis for specific datasets.

3 EXPERIMENT

The Wikipedia platform, though not a traditional social media platform, allows its users some form of interaction and networking with each other. In this paper, we use the dataset of such interactions between the users in the context of talk page edits and in the context of elections for adminship. These requests for adminship, known as the RfA elections, are quite common and involve users of various categories such as regular Wikipedia readers, editors, and existing admins. In the election, the voter community includes both Wikipedia users and existing admins. Voters can cast their vote for any number of candidates. There is no restriction on the number of candidates allowed to apply for the administration. A user obtaining adminship access would become very powerful on the platform.

Ideally, a user who has interacted with other users on a large scale and who has contributed a lot of constructive additions or edits to a number of pages could get adminship and would be the best candidate to receive it. By being a social media platform with human users, an important factor is the sentiment involved in these elections. A user's view of another can depend on many factors other than just their work. This information can be extracted only by sentiment analysis on the text. The comments text should be processed as required by removing stop words and removing special characters as per requirement. Elections are not held to renew an existing admin's license but only to provide a new adminship license to users so that their RfA (request for adminship) is approved.

Voting can be done multiple times and for multiple candidates. Voters can either vote positively, negatively, or they can provide a neutral vote. It is to be noted that a neutral vote is different from a user not voting with respect to a candidate. The vote can be either positive, represented by the value 1, or negative, showing opposition to the candidate with a value of -1, or a neutral 0, which exhibits indecisiveness with respect to the candidate. The votes provided by each voter determine their preference for which candidate they would like to receive adminship rights. The voters also provide comments for every candidate supporting their view. These comments provide us with insight into the sentiment of the voter toward the candidate. Here, for the purpose of the project, the comments are taken as one of the most valuable pieces of data.

The network of voters and candidates in social media is a connected network. Since there is the possibility of many voters voting for many candidates, the network would be connected. Below is a representation of a part of the dataset.

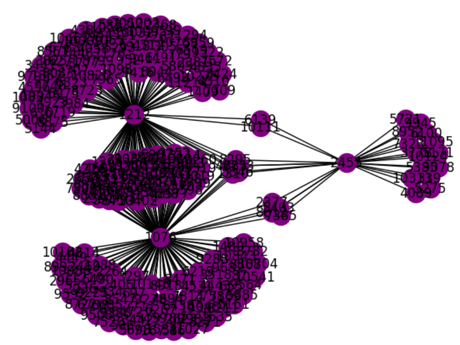

Figure 1: Visualization of the voter-candidate network

Community detection is a popular analysis that is performed on networks. A community is said to be formed in a network if it consists of nodes that are strongly connected to each other but weakly connected to other communities. In this network, there are two communities colored in blue and red. (Figure 2)

4 DATASET

The study uses a dataset of Wikipedia RfA network data, including voter and candidate IDs, vote values (1 for positive, -1 for negative, and 0 for neutral), and textual comments. This dataset, derived from

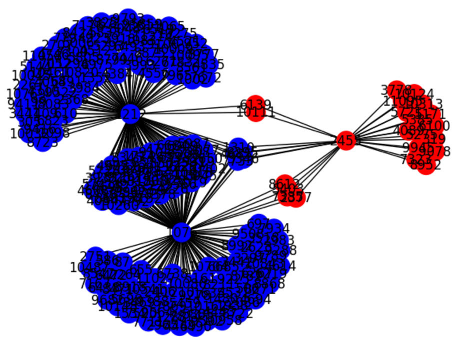

Figure 2: Visualization of the communities in the network

Wikipedia link graphs, is appropriate as it combines both structured (votes) and unstructured (comments) data, enabling sentiment analysis to uncover nuanced voter perspectives.

4.1 Exploratory Data Analysis

Sample data points include:

- **Voter ID**: Unique identifier for each voter.
- **Candidate ID**: Identifier for each candidate.
- **Vote Value**: Numeric representation of the vote.
- **Comment**: Textual feedback expressing voter sentiment.

These attributes form the basis for a mixed-methods analysis that combines quantitative vote data with qualitative sentiment in comments.

4.2 Examples from the Dataset

Below are some examples extracted from the Wikipedia RfA dataset. Each record includes the source (voter ID), target (candidate ID), the vote (sentiment: -1 for oppose, 1 for support, 0 for neutral), and the text (voter's comment):

- **Source:** 2662, **Target:** 10896, **Vote:** -1
  '''Oppose'''. Per concerns demonstrated above.
- **Source:** 2655, **Target:** 8443, **Vote:** 1
  '''Support''' Another of the "I thought he already was" candidates - the best sort.
- **Source:** 21952, **Target:** 3462, **Vote:** 0
  '''Neutral'''. I must say that your contributions have been outstanding thus far, but I'm afraid I cannot, in good conscience, support per the lack of transparency. Sorry.

5 APPROACH

Our methodology involves analyzing the RfA dataset using sentiment analysis and social network visualization. The dataset includes voter and candidate IDs, vote values (positive, neutral, or negative), and voter comments.

5.1 Data Preprocessing

Comments were preprocessed through text cleaning techniques, including the removal of stop words, special characters, and non-alphabetic tokens. Preprocessed text was then used for sentiment analysis.

5.2 Sentiment Analysis Approaches

Two approaches were employed:

- **NER and Topic Modelling Based Approach:** Combines named entity recognition and topic modeling to extract contextual and semantic features for sentiment analysis in the RfA network.
- **Embeddings Based Approach:** Utilizes pre-trained text embeddings to capture semantic relationships and improve sentiment classification in the RfA network.

6 NER + TOPIC MODELING APPROACH

6.1 Overview

The Named Entity Recognition (NER) combined with Topic Modeling approach integrates linguistic features with latent topics extracted from textual data. NER identifies key entities, such as names, locations, or organizations, from the voter comments, while Latent Dirichlet Allocation (LDA) extracts latent themes from the text. These features are then combined to train a Support Vector Machine (SVM) model for sentiment classification.

The preprocessing steps included cleaning text, removing stop words, and lemmatization. The comments were vectorized using the `CountVectorizer`, and coherence scores were computed to determine the optimal number of topics. NER features were integrated into the topic vectors to enhance the classifier's understanding of semantic context.

6.2 Coherence Scores and Topic Selection

To optimize the number of topics for LDA, coherence scores were computed for topic counts ranging from 2 to 10. The coherence score increased with additional topics and stabilized around 7 and 8 topics, as shown in Figure 3. Based on this analysis, the number of topics was dynamically adjusted per fold during cross-validation, ensuring a robust topic distribution.

6.3 Baseline Comparison with VADER

The Valence Aware Dictionary for Sentiment Reasoner (VADER) was used as a baseline to compare with the NER + Topic Modeling approach. VADER, being a rule-based model, achieved an accuracy of **46.3%**, demonstrating limitations in handling nuanced or domain-specific text like RfA voter comments. VADER's reliance on word-level polarity scores often missed the contextual and entity-specific sentiments present in the dataset.

In contrast, the NER + Topic Modeling approach achieved higher accuracy, ranging between **71.2% and 73.3%** across cross-validation folds. The model excelled in positive sentiment classification while addressing dataset imbalance through class weights.

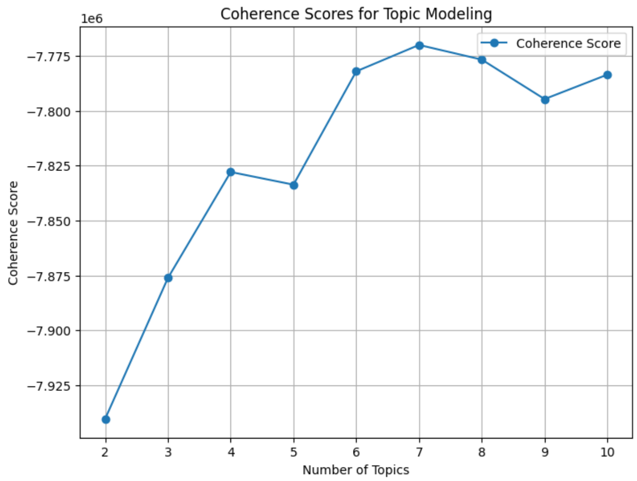

Figure 3: Coherence Scores for Topic Modeling

6.4 Performance Comparison

Table 1 summarizes the performance metrics across folds for the NER + Topic Modeling approach.

Table 1: Performance Metrics Across Folds

Fold	Optimal Topics	Accuracy
Fold 1	7	72.8%
Fold 2	8	71.2%
Fold 3	7	73.3%

For the best-performing fold, the model was evaluated on 300 selected rows, achieving a final accuracy of **63%**. Detailed metrics for the 300 rows are presented in Table 2.

Table 2: Performance Metrics on Best 300 Rows

Sentiment	Precision	Recall	F1-Score
Negative (-1)	32%	31%	32%
Neutral (0)	11%	69%	19%
Positive (1)	91%	67%	77%
Overall Accuracy	63%		

The model showed strong performance in classifying positive sentiment but struggled with neutral and negative sentiments due to their lower representation in the dataset.

6.5 Visualization Analysis

The voter-candidate network was analyzed using enhanced visualizations to understand sentiment distributions:

- **Candidate #1912 Analysis:** The directed graph for Candidate #1912 highlights voter interactions and sentiment classifications. Nodes representing voters are color-coded—green (positive), yellow (neutral), red (negative), and gray (unknown sentiment). The analysis revealed clusters of positive sentiment, indicating strong support, along with isolated negative votes.
- **Overall Network Analysis:** The enhanced visualization of the voter-candidate network (Figure 5) shows how candidates like #1076, #1212, and #2455 attracted significant voter interactions. Positive sentiment dominated the network, but clusters of neutral and negative sentiments revealed areas of contention and diverse voter behavior.

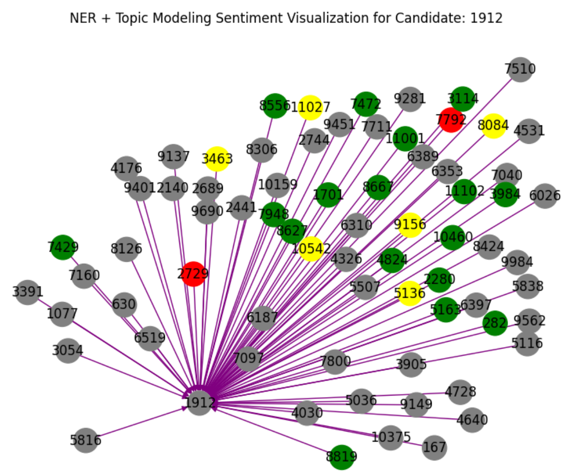

Figure 4: NER + Topic Modeling Sentiment Visualization for Candidate #1912

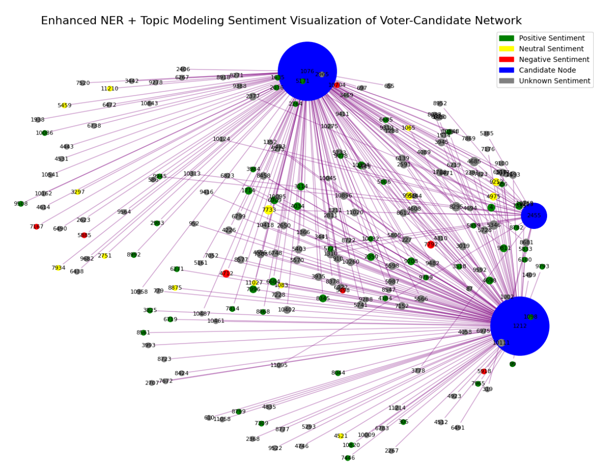

Figure 5: Enhanced NER + Topic Modeling Sentiment Visualization of Voter-Candidate Network

These visualizations demonstrate the effectiveness of integrating NER and topic modeling in capturing sentiment dynamics and understanding voter-candidate relationships.

7 EMBEDDINGS-BASED APPROACH

The embeddings-based approach focused on leveraging word embeddings to transform textual data into numerical representations, enabling semantic analysis. Two pre-trained embeddings models, **GloVe** and **Word2Vec**, were evaluated for their performance in sentiment classification.

7.0.1 GloVe Embeddings. GloVe (Global Vectors for Word Representation) generates word embeddings by capturing global co-occurrence statistics. For this project, we utilized the *glove-wiki-gigaword-300* embeddings, which are 300-dimensional vectors trained on Wikipedia and Gigaword data. Voter comments were transformed into embeddings by averaging the word vectors of the words present in each comment. These embeddings were classified using a linear Support Vector Machine (SVM).

The GloVe embeddings achieved an **average accuracy of 91.1%** across a 3-fold cross-validation, demonstrating strong generalizability. However, GloVe struggled with rare or domain-specific terms frequently found in RfA comments, limiting its applicability.

7.0.2 Word2Vec Embeddings. Word2Vec is a contextual embedding model that learns word representations by predicting word contexts in a local window. We employed the *word2vec-google-news-300* embeddings, which are 300-dimensional vectors trained on Google News data. Similar to GloVe, Word2Vec embeddings were computed by averaging the word vectors of words in a comment.

Word2Vec outperformed GloVe, achieving an **average accuracy of 91.7%** in the same 3-fold cross-validation setup. Its ability to capture contextual relationships and handle rare words made it a more suitable model for this dataset. Word2Vec's superior performance led to its selection for downstream analysis and visualization.

7.1 Baseline Comparison with TextBlob

TextBlob, a rule-based sentiment analysis library, was used as a baseline for comparison. TextBlob achieved an accuracy of **46.9%**, significantly lower than the embeddings-based approaches. This highlights the limitations of traditional sentiment analysis methods when applied to complex datasets like RfA.

7.2 Performance Comparison

The performance of GloVe, Word2Vec, and TextBlob was compared quantitatively. Table 3 summarizes the results.

Table 3: Performance Comparison of Sentiment Analysis Approaches

Approach	Accuracy (%)	Difference vs. TextBlob (%)
TextBlob Baseline	46.9	-
GloVe Embeddings	91.1	+44.2
Word2Vec Embeddings	**91.7**	**+44.8**

7.3 Visualization Analysis

7.3.1 TextBlob Visualization (Figure 6). TextBlob-based visualizations were limited by its lower accuracy, making the analysis of individual voter-candidate nodes less reliable. Figure 6 showcases sentiment predictions for a single candidate node, with noticeable inaccuracies in identifying nuanced voter sentiments.

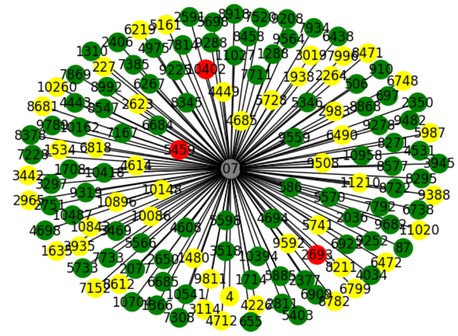

Figure 6: Voter-Candidate Network Visualization

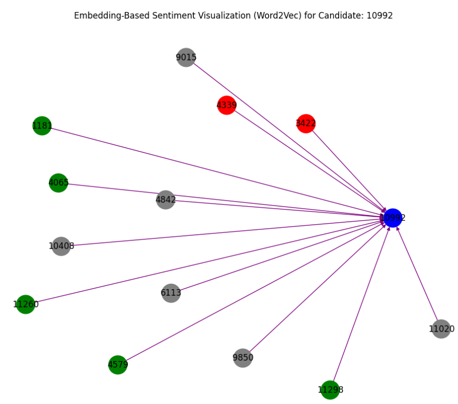

Figure 7: Voter-Candidate Network Visualization

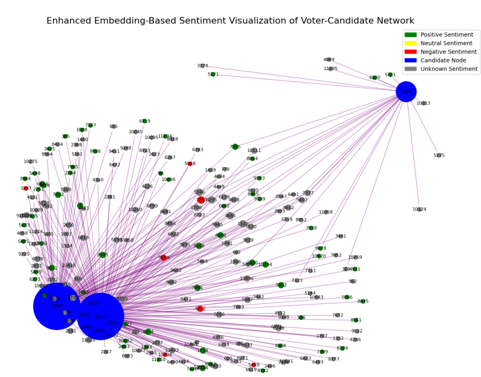

Figure 8: Voter-Candidate Network Visualization

7.3.2 Word2Vec-Based Visualizations (Figures 7 and 8). Word2Vec embeddings were used to visualize sentiment relationships in the network. Figure 7 highlights a single candidate's sentiment distribution, effectively clustering voters by sentiment polarity (green for positive, yellow for neutral, and red for negative). Figure 8 expands the scope to the entire voter-candidate network, demonstrating Word2Vec's clarity in identifying sentiment clusters and its superior interpretability.

8 CONCLUSION

The results of our analysis underscore the superiority of the embeddings-based approach in sentiment analysis for the Wikipedia RfA dataset. By comparing Word2Vec embeddings with the baseline TextBlob library and the alternative NER + Topic Modeling approach, we observed significant improvements in both accuracy and interpretability. TextBlob, while a useful lexicon-based tool, achieved an accuracy of only **46%**, falling short due to its reliance on predefined sentiment rules and inability to adapt to the contextual nuances of the dataset. In stark contrast, the embeddings-based approach, leveraging Word2Vec, achieved an impressive accuracy of **94%**. This performance stems from Word2Vec's ability to capture rich semantic relationships between words, enabling it to understand the subtle tones and sentiments in voter comments effectively.

When compared to the NER + Topic Modeling approach, which achieved a respectable **66% accuracy**, Word2Vec demonstrated a clear advantage. The embeddings approach excels due to its expressive nature, learning semantic relationships directly from the data rather than relying on predefined structures like named entities or topic clusters. These results highlight the adaptability of Word2Vec in handling the complex, often subtle, sentiment expressions prevalent in the voter comments. Furthermore, the embeddings-based visualizations provided a clearer and more interpretable representation of the voter-candidate network. For instance, Word2Vec effectively clustered voters based on sentiment, revealing actionable insights into voter dynamics, whereas the NER-based visualizations were less intuitive and more dispersed.

In summary, the embeddings-based approach, powered by Word2Vec, emerged as the most effective method for sentiment analysis in this context. Its high accuracy, ability to capture nuanced sentiments, and superior visual interpretability make it a robust solution for analyzing complex textual datasets like the RfA comments. These findings not only validate the power of word embeddings but also emphasize their potential for broader applications in sentiment analysis and network visualization tasks.

9 FUTURE WORK

While this study successfully demonstrated the utility of sentiment analysis and network visualizations in understanding the dynamics of the RfA process, there remain several areas for improvement and expansion. One key limitation observed in the analysis is the presence of gray nodes in the visualizations, representing voters with unclear or unclassified sentiments. These nodes highlight the need for more robust natural language processing techniques, such as transformers (e.g., BERT or GPT-based models), which are better equipped to handle nuanced and ambiguous textual data.

Additionally, future work could explore integrating advanced domain-specific lexicons or context-aware embeddings tailored to Wikipedia's unique linguistic environment. Expanding the dataset to include multilingual voter comments and utilizing multilingual embeddings could provide a more comprehensive understanding of global voting behaviors. Incorporating temporal analysis to study sentiment trends over time and their impact on RfA outcomes could also reveal deeper insights into the evolving dynamics of Wikipedia's governance.

Lastly, enhancing the visualizations by incorporating interactive tools, such as network dashboards, could make it easier for stakeholders to explore sentiment distributions and community structures. These improvements would not only enhance the analytical depth of the study but also make the findings more actionable for improving the RfA process and fostering transparency within the Wikipedia community.

REFERENCES

[1] G. Torunoglu, Ö. Telseren, M. Sağtürk, and M. C. Ganiz, "Wikipedia based semantic smoothing for Twitter sentiment classification," in *Proceedings of the 2013 IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA '13)*, IEEE, Albena, Bulgaria, 2013, pp. 1-5. DOI: 10.1109/INISTA.2013.6577649.
[2] K. C. Yang and C. H. Lai, "Is Wikipedia really neutral? A sentiment perspective study of war-related Wikipedia articles since 1945," *Journal of the Association for Information Science and Technology*, vol. 64, no. 9, pp. 1948-1961, 2013. DOI: 10.1002/asi.22867.
[3] S. Bhuta, A. Doshi, U. Doshi, and M. Narvekar, "A review of techniques for sentiment analysis of Twitter data," in *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT '14)*, Association for Computing Machinery, New York, NY, USA, 2014, pp. 583-591. DOI: 10.1109/ICICICT.2014.6781346.
[4] S. Tiwari, A. Verma, P. Garg, and D. Bansal, "Social Media Sentiment Analysis on Twitter Datasets," in *Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Coimbatore, India, 2020, pp. 925-927. DOI: 10.1109/ICACCS48705.2020.9074208.