Leveraging LLMs for Enhanced Detection of Controversies On Wikipedia

Rudhar Pratap Singh[1], **Aryan Dua**[2] **and Ananya Aakriti**[3]

Abstract -

This report focuses on the application of language models for controversy detection. We establish baselines through a meticulous examination of four key methodologies, mentioned in the baselines section. The scalability of our dataset is also scrutinized, considering the feasibility of its expansion. We explore avenues for potential growth and discuss implications for scaling up the dataset to accommodate diverse applications and increased robustness.

1 Introduction

Controversial content on Wikipedia can lead to differing opinions and affect the platform's credibility as a reliable source of information. There is a need to develop a robust methodology to detect and analyze controversial content on Wikipedia pages. One idea can be to utilize the edit history and textual data of Wikipedia articles. Edit histories may reveal edits of differing opinions. One may gain valuable insights into the emergence and evolution of controversies in collaborative online environments.

2 Baselines

To benchmark the effectiveness of using LLMs, we will assess its performance against existing models that serve as reference points in controversy detection. These baseline models will encompass conventional methods and established techniques commonly used in the field. By contrasting the outcomes of our LLM-based approach with these baseline models, we aim to gain insights into the novel contributions and improvements our proposed methodology offers. This comparative analysis will provide a comprehensive foundation for evaluating the efficacy of LLMs in detecting controversies on Wikipedia, setting the stage for a more in-depth examination of our innovative approach in subsequent sections.

2.1 Logistic Regression on Edit Counts

The edit count serves as a straightforward quantitative metric reflecting the level of activity and dynamism surrounding a particular article. Logistic regression, a statistical modeling technique, was employed to analyze the relationship between the edit count and the likelihood of an article being controversial. By employing this approach, we sought to assess whether the sheer volume of edits could serve as a predictive factor for controversy, offering a more conventional and easily interpretable metric compared to the linguistic complexities addressed by language models. This analysis not only contributes a quantitative perspective to the controversy detection task but also allows for a comparative evaluation against the performance of other baseline models and our proposed LLM-based methodology.

The maximum accuracy achieved by this approach was 75.9%.

2.2 Naive Bayes' Classification on article content

Leveraging this classifier involved training the model on a labeled dataset of Wikipedia articles, where each article was annotated as controversial or non-controversial. The classifier utilizes the probabilistic relationship between words and the likelihood of an article being controversial or not. By employing this approach, we aimed to gauge the performance of a traditional and well-established classification technique before delving into the intricate nuances introduced by Large Language Models (LLMs).

The maximum accuracy obtained from this model was 86.2%.

Below are the wordclouds obtained for controversial articles and non-controversial articles:

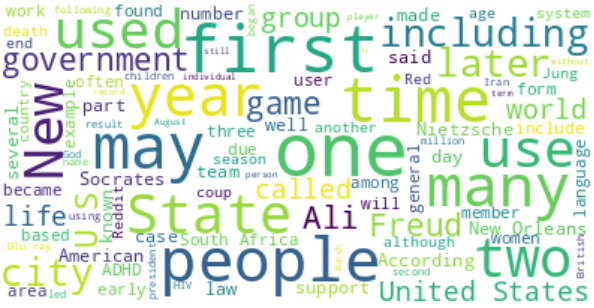

Figure 1. Wordcloud of controversial articles

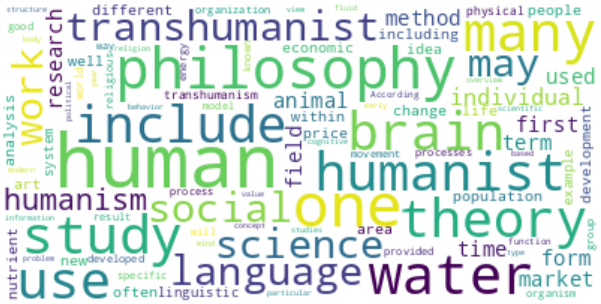

Figure 2. Wordcloud of non-controversial articles

2.3 Logistic Regression on edit count and article content

We extended our investigation by combining logistic regression with two distinct features: the edit count and content-based features derived from CountVectorizer. The edit count encapsulates the dynamic evolution of an article, serving as a numerical indicator of its activity. Simultaneously, CountVectorizer was employed to transform the textual content of each article into a numerical representation, capturing the frequency distribution of words. By integrating these features, we aimed to leverage both quantitative metrics and linguistic patterns to enhance the predictive power of our model. Logistic regression was then applied to analyze the combined feature set, allowing us to discern the nuanced relationship between edit frequency, textual content, and the likelihood of an article being classified as controversial.

The maximum accuracy achieved by this approach was 93.1%.

2.4 Using Controversiality Scores

We implemented the technique mentioned in: "The most controversial topics in Wikipedia: A multilingual and geographical analysis by Taha Yasseri, Anselm Spoerri, Mark Graham, and János Kertész". They quantify the controversiality of an article based on its editorial history, by focusing on "reverts", i.e. when an editor undoes another editor's edit completely and brings it to the version exactly the same as the version before the last version.

To detect reverts, they first assign a hash code to each revision of the article and then by comparing the hash codes, detect when two versions in the history line are exactly the same. In this case, the latest edit (leading to the second identical revision) is marked as a revert, and a pair of editors, namely a reverting and a reverted one, are recognized. A "mutual revert" is recognized if a pair of editors (x, y) is observed once with x and once with y as the reverter. The weight of an editor x is defined as the number of edits N performed by him or her, and the weight of a mutually reverting pair is defined as the minimum of the weights of the two editors. The controversiality M of an article is defined by summing the weights of all mutually reverting editor pairs, excluding the topmost pair, and multiplying this number by the total number of editors E involved in the article. This results in the following formula:

$$M = E \sum_{\text{all mutual reverts}} min(N^{\mathrm{d}}, N^{\mathrm{r}})$$

where N r/d is the number of edits for the article committed by reverting/reverted editor. The sum is taken over mutual reverts rather than single reverts because reverting is very much part of the normal workflow, especially for defending articles from vandalism. The minimum of the two weights is used because conflicts between two senior editors contributing more to controversiality than conflicts between a junior and a senior editor, or between two junior editors. And finally the topmost reverting pair is excluded to avoid overestimating the editorial war dominated by a personal fight between two single editors. The explained measure can be easily calculated for each article, irrespective of the language, size, and length of its history.

Using this formula to calculate the controversiality score, we find an appropriate classification threshold and the maximum accuracy we can report using this technique is 77.4%

3 LLM-based approaches

LLM-based approaches are ideal for this scenario because these models excel in understanding contextual nuances and bidirectional language patterns, providing a robust framework for capturing the complexity inherent in controversial topics. Their pre-trained representations, derived from extensive and diverse text data, enable a broad understanding of language, while transfer learning allows for fine-tuning on specific controversial language characteristics.

LLMs are adept at handling ambiguity and multiple perspectives, crucial in the context of controversies. The rich feature representations they generate encompass syntactic, semantic, and contextual information, facilitating nuanced predictions. Moreover, their adaptability allows for fine-tuning on datasets related to controversial topics, tailoring the model to the specific language nuances found

in Wikipedia articles. While the choice of the optimal model depends on various factors, including dataset size and computational resources, LLMs offer a powerful and versatile tool for effectively classifying controversial content.

3.1 0-Shot Model: BART NLI Model

We employed pre-trained Natural Language Inference (NLI) models as 0-shot sequence classifiers to detect controversies in Wikipedia articles. This section elaborates on the methodology used for classification without explicit training data.

3.1.1 Classification Method

In the context of this model, sequences were treated as premises, with hypotheses formed for candidate labels. This approach made it feasible to perform classification without the need for specific training data. By utilizing the NLI models, the process involved mapping the sequences as premises and forming hypotheses to predict candidate labels.

3.1.2 Probability Transformation

The NLI model's entailment and contradiction probabilities played a pivotal role in the classification process. These probabilities were transformed into label probabilities, simplifying the classification task without requiring dedicated training data. This transformation allowed for the conversion of the inherent entailment and contradiction probabilities into probabilities associated with the specific labels under consideration, facilitating the identification of controversial topics in Wikipedia.

This methodology leveraged the inherent capabilities of the BART NLI model, enabling effective detection of controversies without the necessity of extensive labeled datasets for training.

3.1.3 Limitations

1. **Pre-training**
   Natural Language Inference (NLI) models, such as the BART NLI model, while proficient in various natural language understanding tasks, lack specialization in identifying controversial content. These models are not inherently tailored or fine-tuned explicitly for sensitivity towards controversial topics, limiting their effectiveness in controversy detection.

2. **Variable Controversy Detection**
   The effectiveness of NLI models in detecting controversies exhibits variability based on the available training data and the complexity of the task. Pretrained models, even when using 0-shot approaches, might not consistently capture controversies without specific fine-tuning for this purpose. Fine-tuned models, which are adjusted or trained specifically for detecting controversies, tend to provide more consistent and accurate results.

3. **Relatively Low Accuracy**
   Due to the inherent nature of NLI models not being specialized in controversy detection, the overall accuracy in identifying controversies using 0-shot methods might be relatively low. Without fine-tuning or specialized training, these models may struggle to capture nuanced, context-dependent controversial topics present in Wikipedia articles.

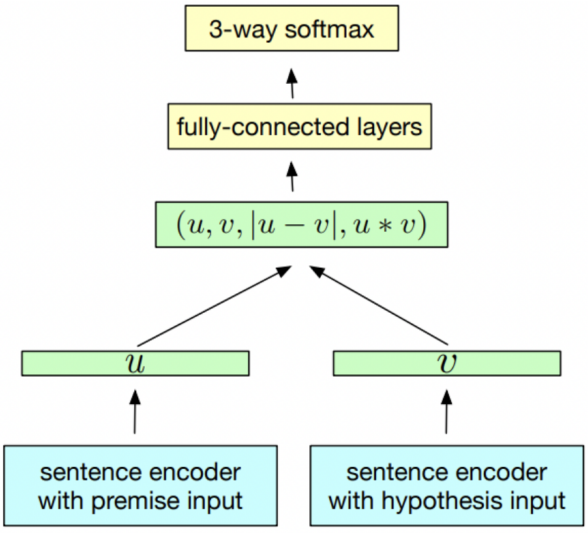

Figure 3. Generic NLI training scheme

3.1.4 Accuracy

The accuracy achieved was 65%

3.2 Few-Shot Model: RoBERTa Base Model

RoBERTa, a transformer-based model, undergoes pretraining on an extensive English text corpus using self-supervised learning techniques, eliminating the need for human labeling. The model's characteristics and pretraining methodology are outlined as follows:

3.2.1 Masked Language Modeling

During the pretraining phase, RoBERTa employs Masked Language Modeling (MLM) by randomly masking 15% of the words within sentences. It then predicts these masked words, facilitating the model's ability to comprehend contextual information from both forward and backward directions in the text.

3.2.2 Bidirectional Understanding

Unlike conventional models, RoBERTa excels in learning bidirectional sentence representations. This approach allows the model to capture richer contextual information by understanding relationships between words and phrases in both directions within the text.

3.2.3 Feature Extraction

RoBERTa, owing to its robust pretraining methodology, serves as an effective feature extractor. The representations learned during pretraining can be utilized for various downstream tasks, significantly enhancing performance in classification and other Natural Language Processing (NLP) applications.

3.2.4 Accuracy

The accuracy achieved was 44%

3.3 Integrating Edit counts

1. Along with topic name, also add edit count of articles to the prompts

2. The model gains insight into the significance and context of each topic.

3.3.1 Accuracy

The accuracy achieved was 80%

3.4 Summarizer Model

We employed the Pegasus Model as a summarizer, utilizing its capabilities for generating concise prompts. The key methodology and components of the Summarizer Model are detailed below:

3.4.1 Pegasus Model Implementation

The chosen Summarizer Model, Pegasus, was integrated into the research framework for its proficient summarization capabilities. It facilitated the creation of concise prompts, condensing complex information into shorter representations.

3.4.2 Prompt Components

To formulate the prompts, we incorporated essential elements such as the topic name, the number of edits made to the article, and the article summary. These components were strategically combined to create informative yet concise prompts.

3.4.3 Clarity and Relevance in Generated Content

The utilization of the Summarizer Model ensured the clarity and relevance of the generated content. By synthesizing crucial information into succinct prompts, the model contributed to the precision and informativeness of the generated outputs.

3.4.4 Accuracy

The accuracy achieved was 81%

3.5 Subtopics and Examples Model

We implemented a categorization strategy involving nine distinct subtopics using 0-Shot Prompting. This categorization methodology facilitated the organization of articles into specific thematic clusters. The key components and approach taken for subtopic categorization are detailed as follows:

3.5.1 0-Shot Prompting for Subtopic Categorization

The subtopic categorization process relied on 0-Shot Prompting techniques, which allowed for categorizing articles without the need for explicit training data. This approach involved formulating prompts that encompassed both controversial and non-controversial articles within each subtopic.

3.5.2 Prompts Connecting Articles within Subtopics

For every article, prompts were strategically constructed to establish connections to related articles within the same subtopic. These prompts served the purpose of contextualizing each article within its respective thematic cluster, enhancing coherence and thematic relevance.

The 9 Subtopics:

1. Social Sciences and Society
2. History and Events
3. Religion and Spirituality
4. Health and Fitness
5. Art and Culture
6. People and Self
7. Natural Sciences and Nature
8. Technology and Applied Sciences
9. Philosophy and Thinking

3.5.3 Accuracy

The accuracy achieved was 87%

3.6 Text Similarity Model

We used a text similarity model which was employed to assess the similarity between articles. This process involved identifying articles with higher text similarity to a given input article, forming the basis for selecting prompts. The following details outline the methodology and key components used for gauging text similarity and selecting appropriate prompts:

3.6.1 Utilizing BERT for Similarity Assessment

The BERT (Bidirectional Encoder Representations from Transformers) model was utilized for assessing text similarity. We used the 'sentence-transformers' Library. This library extends the Hugging Face Transformers for sentence embeddings. It provides pre-trained models that can convert sentences into dense vectors (embeddings). These embeddings can be used for various NLP tasks, including finding similar sentences.

3.6.2 Cosine Similarity Calculation

To quantify text similarity, cosine similarity was calculated between the article representations obtained from BERT. This calculation method facilitated the identification of the top 5 articles with the highest similarity to the input article, serving as examples to the prompts (examples are required as we are using few-shot prompting).

3.6.3 Few-Shot Prompting

The top 5 most similar examples that were generated by the BERT model, are now fed to a LLM API, which then uses these 5 examples to predict whether a given article is controversial or not. This is known as Few-Shot Prompting.

3.6.4 Accuracy

The accuracy achieved was 84%

4 Conclusions and Future Work

Certain results derived from the baseline models have demonstrated higher accuracy in comparison to our Large Language Model (LLM)-based approaches. It is important to note, however, that these reported accuracies stem from a notably small test set, comprising only 29 examples. Consequently, the interpretation of these results should be approached with caution, given the limited scale of the evaluation dataset. The small sample size may not fully capture the robustness and generalizability of the models.

Ideally, our model should perform better as it incorporates all the features of the baseline models. As we proceed with our research, expanding the test set size and incorporating diverse examples will be imperative for a more comprehensive and reliable assessment of the comparative performance between the baseline models and our LLM-based methodologies. This caveat underscores the necessity for continued experimentation and refinement to draw more conclusive insights about the efficacy of our proposed approach.

5 References

1. 'The most controversial topics in Wikipedia: A multilingual and geographical analysis' - Taha Yasseri, Anselm Spoerri, Mark Graham, and János Kertész

2. 'Learning to Retrieve In-Context Examples for Large Language Models' - Liang Wang, Nan Yang, Furu Wei