Controversy In Wikipedia Articles

Xingyu Jiang, Xiangchen Zhao, Hengyu Liu

Halıcıoğlu Data Science Institute University of California, San Diego

Abstract

There are “wars” going on every day online, but instead of cities, they are defending their options and perspects. This phenomenon is especially common on the Wikipedia platform where users are free to edit others' revisions. In fact, there are “about 12% of discussions are devoted to reverts and vandalism, suggesting that the WP development process is highly contentious.” (Robert 1) As Wikipedia has become a trusted source of information and knowledge which is freely accessible, It is important to investigate how editors collaborate and controvert each other in such a platform. This paper will discuss a new method of measuring controvisality in Wikipedia articles. We have found out that controversiality is highly related to the number of revert edits, the sentiment level among one article comments, and the view counts of that article. Thus we developed a weighted sum formula, which combines those three factors to accurately measure the controversy level within articles in Wikipedia.

Keywords: Controversy, Wikipedia articles, Sentiment, View counts, M-score

Introduction

Wikipedia is now a trusted source of information and knowledge which is freely accessible. Since it has become such an important resource, its information quality needs to be monitored for providing more accurate sources to readers. Thus, it is important to look at how Wikipedia information is generated and how people make changes to Wikipedia information. Investigating this allows us to gage the trustworthiness of Wikipedia articles and allows for the discovery of methods to collaborate more efficiently.

Our project is built on a previous paper called *Edit wars in Wikipedia* which investigates a similar topic. The paper believes that analyzing revisions is a good way to build an understanding of an article's controversy since a reversion erases the work of a user and mutually doing it between editors is a good indicator that they disagree enough not to be able to work this problem out through other channels. Their measurement of controversy, known as the "M-statistic", is calculated in the following way: for every revision, check to see if there is a revert. For each revert, they take the min amount of user edits out of the following 2 categories: 1) the person that reverted the article 2) the person that was reverted. This min amount is then summed across all reverts (excluding the top 2 users). This amount is multiplied by the number of users. However, we found out that M-statistics fails to reflect the controversy level accurately within non-popular articles, which consists nearly 95% of all the articles in Wikipedia. In fact, we found out 99% of articles have a 0 M statistic, which means either there is no evidence of controversy at all or the M-score cannot detect the controversy in those articles. According to our analysis, the shortcomes of M statistic are as follows:

First, M statistics does not look deeper into each revert. For example, a revert could be made because the reverted edit is wrong, or because of vandalism, or even could be an outcome by bots' routine work. M statistic will tread a revert equally regardless of its reason. Secondly, by only using the number of mutual reverts and the number of people who make mutual reverts as a way to evaluate the controversy, it misses a lot of parts in Wikipedia articles that may also contain some contents that can directly show you this article has a high level of controversy, such as some negative revision content, pageview counts for articles and etc.

Thus it inspires us to generate a more accurate weighted sum formula by including M statistic, sentiment analysis of comments, article pageview counts to evaluate the controversy of Wikipedia articles. By including these additional factors, we wish our final formula could perform better within non-popular articles (where M statistic is 0), and largely maintain the accuracy of controversy level within popular articles.

Background

Wikipedia articles are a great free accessible data source that people use to find information they need. And this feature also lets Wikipedia articles become a source for letting people make research. For example, there is research that is trying to explore areas of collaboration that result in quality content, Shane Greenstein and Feng Zhu, authors of "*Do Experts or Crowd- based Models Produce More Bias?*", looked at the differences between "expert" opinions and Wikipedia articles bias, found that articles with more revisions tended to be more neutral; however, it's effect on bias is not strong and many articles do not undergo that many revisions. Also, there is another research that is related to the Wikipedia article and the result from Shane

Greenstein and Feng Zhu's paper. In Kittur and Kraunt's paper, "Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination", they used people's evaluations of quality (through mechanical turks) and Wikipedia's own article assessment project as a measurement, and they found that more editors working with certain amounts of implicit coordination led to the most quality results (however, just adding more editors did not necessarily mean an increase in quality). We also think it is pretty interesting to make analysis on Wikipedia articles, which lets us generate the idea to analyze the controversy of Wikipedia articles.

Datasets and Methods

Raw Edit History Dump:

The first dataset is an edit with comments dump file, we get this dataset from the raw XML file. The data looks like this:

	commentor	date	comment	Revision Time
0	Vodex	2004-03-27 17:54:55	NaN	0
1	Vodex	2004-04-05 18:48:35	expanded article - need more examples. How to ...	1
2	Vodex	2004-04-08 09:53:14	ref. Kirt Angle	2
3	Dale Arnett	2004-04-13 06:02:24	=Examples of angles=	0
4	Dale Arnett	2004-05-07 19:09:32	=Examples of angles=	1
...	...	...	...	...
65105	Pengo	2006-01-02 00:22:32	pPpPp	3
65106	Pengo	2006-01-06 01:41:10	NaN	4
65107	Pengo	2006-01-06 01:45:50	NaN	5
65108	Pengo	2006-01-06 01:48:42	NaN	6
65109	Pengo	2006-01-06 02:09:20	redundant article	7

Figure 1: detailed information about edit with comments dump file

In this dataset, it contains the commenter names, titles of articles that they make comments on, and the time for them to make any comments in Wikipedia articles. Besides those columns, it also contains a column named "*comment*", which is the content of those commenters' comments. For this dataset, we mainly want to get the *comment* column, which we will use in our sentiment analysis part to analyze the sentiment score for each comment. Later on, we will discuss how to use comments to detect the level of controversy within one article. Overall we believe that sentiment score can be an important factor in our weighted sum formula to calculate the extent of controversy for Wikipedia articles. Also, we generate both columns named as "*commentator*" and "*time*" for merging this dataset with the English Light Dump file.

Note that the entire edit history for all Wikipedia articles is huge and beyond our parse ability. This project will only use 2600 articles and their entire edit history (around 30GB). Our data processing pipeline will automatically download the selected bz2 file and read the xml line by line to extract the necessary data fields (Contributor name, comments, etc.).

English Light Dump:

The second dataset is English light Dump file, we get this dataset from WikiWarMonitor. The dataset looks like this:

	date	linelabel	revised_time	commentor	article
0	2011-10-28 12:51:21	0	2	MystBot	Herbert_Art_Gallery_and_Museum
1	2011-10-28 12:28:07	0	1	Rock_drum	Herbert_Art_Gallery_and_Museum
2	2011-10-28 08:32:39	0	7	DJDunsie	Waiting_to_Exhale
3	2011-10-28 08:32:21	0	6	DJDunsie	Waiting_to_Exhale
4	2011-10-28 08:31:56	0	5	DJDunsie	Waiting_to_Exhale
...	...	...	...	...	...
2002390	2004-04-10 13:46:34	0	5	213.164.241.16	August
2002391	2004-04-10 13:14:14	0	4	Webkid	August
2002392	2004-03-08 17:51:16	0	3	HasharBot	August
2002393	2003-12-19 23:05:12	0	2	HasharBot	August
2002394	2003-03-27 11:02:15	0	1	Ams80	August

Figure 2: detailed information about English Light Dump dataset

This dataset is generated to calculate M statistic which contains *"date", "linelabel", "revised_time", "commentor", "article"* columns. For the "*linelabel"* column (we will call it "*revert*" later), it shows whether some parts of the article are reverted by others. And for the *revised_time* column, it shows the number of revisions for the article. We use this dataframe to merge with the edit with comment dataset and get a dataframe that contains all of their information. Also, we use this dataframe to get an *M-score* for each article, because we think although the *M-score* is not so accurate, it is still an important part of our weighted sum formula.

Pageview API:

The third dataset we are using is a dataset generated from the Pageview API. The dataset looks like the following figure:

	title	views
0	Ôdate,_Akita	449.0
1	Ôdate	422.0
2	Yōjirō_Ishizaka	10135.0
3	Yixin,_1st_Prince_Gong	1941.0
4	Xianfeng	10262.0
...	...	...
74	aharoff_Basil	NaN
75	aleucus	NaN
76	awditu	NaN
77	awditu_of_Ethiopia	NaN
78	wischenzug	NaN

Figure 3: detailed information about Pageview dataset

This dataset only contains two columns: For the *"title"* column, it contains the name for each Wikipedia article. The *"view"* column is generated by "title" columns by getting the raw description number of views from the article's start date to January 1st, 2021 from the Pageview API. The reason for us to generate such a dataset is we think pageview counts can be related with the controversy of Wikipedia articles. In other words, we think if an article has large pageview counts, it should have a higher level of controversy.

Sentiment Analysis on Comments

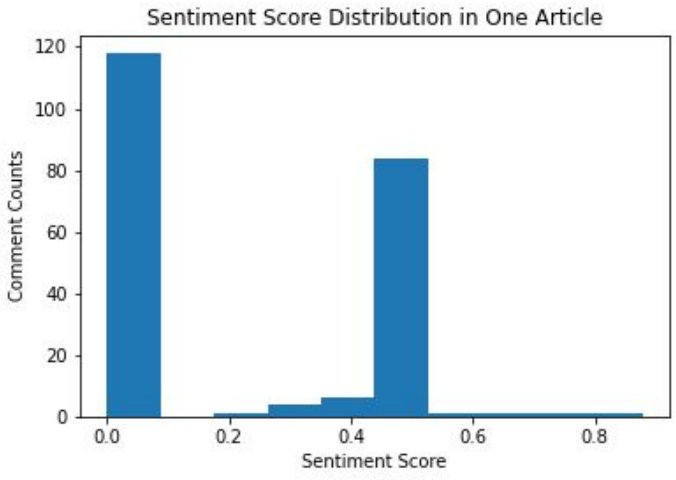
When we examine the entire edit history of one article, We noticed every time an editor gives a revision, he may leave a comment on what is being revised, which is also a perfect place to defend his edit or attack others (figure 4). In order to get the general sentiment statistic for a certain article, we need to run sentiment analysis on every edit comment of one article. Then we

hypothesize that: Bringing in sentiment score could balance M's performance in lower M score range.

```
<page>
  <title>Talk:Page title</title>
  <revision>
    <timestamp>2001-01-15T14:03:00Z</timestamp>
    <contributor><ip>10.0.0.2</ip></contributor>
    <comment>hey</comment>
    <text>WHYD YOU LOCK PAGE??!!! i was editing that jerk</text>
  </revision>
</page>
```

Figure 4: Example of an edit comment

Here is an example that shows the importance of adding sentiment analysis into our controversy measurement. We had found an article "Wooster, Ohio", who has 0 M statistic, but we still detect heavy sentiment inside its comments (see figure 5). This histograms shows even if its M statistic is 0 (leftmost bar), half of its comments consists of a certain degree of sentiment[1]. That shows some validity of why sentiment analysis should be considered when measuring controversy.



[1] The sentiment score for each comment is within [-1,1]. Here we take the absolute value of each score to measure the intensity level of sentiment.

Figure 5: Sentiment distribution of *Wooster, Ohio*

Next, we apply this analysis with all our available data[2]. We plot the scatter plot (figure 6) where its x-axis is the log scaled M statistic and y-axis is the mean sentiment score. If we ignore the dots in the green circle, we will notice there is a good positive correlation between M and sentiment score, which prove that sentiment score and M are both good reflections of controversy in the range. However, when we look at the dots in the green circle, we found that even if they are low in M score, they still vary in sentiment level. This is why we want to use sentiment scores to compensate for M's performance in low M region.

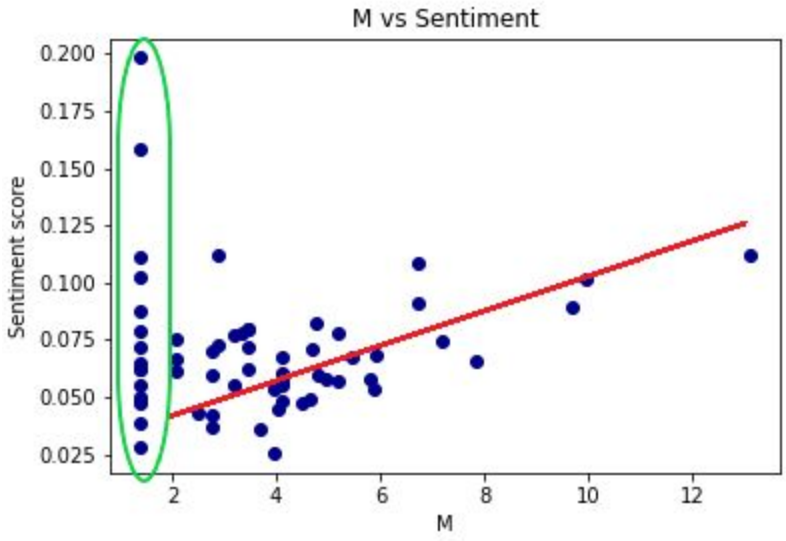


Figure 6: M and Sentiment Score

Note that in this graph, we use the mean absolute sentiment score instead of the total absolute sentiment score of each article to reflect the overall sentiment intensity among all the edits. However, we noticed some articles which should be considered as more controversial (such as the *Republican Party*) have an usual average sentiment. This is due to the large amount of edits

[2] Our dataset consists of a total of 2600 articles with their entire edit history, most of which are related to Ohio State.

as a denominator that makes the mean much smaller. Thus sentiment score will only perform well when that article's corresponding M is small.

Pageview Count and Controversy

The view count of an article normally decides the popularity of that article: The higher view counts means higher popularity. Thus we further hypothesize that may be higher view counts could also mean that the article is more controversial. We decided to use our two existing factors to validate this hypothesis: Sentiment score and M statistic.

Because articles with 0 M statistics consist of 90% of our dataset, we remove all these articles to make the trend more visible. We also log scaled our M statistic and view counts

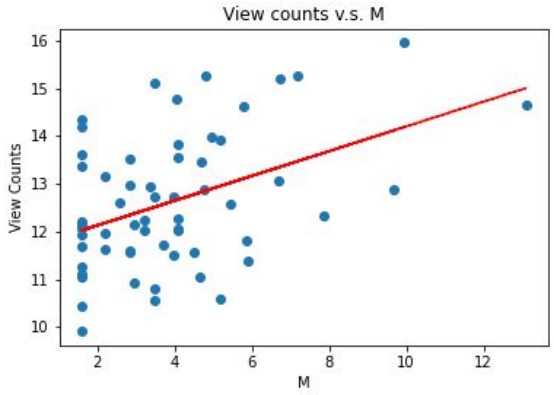


Figure 7: View Counts and M Statistic

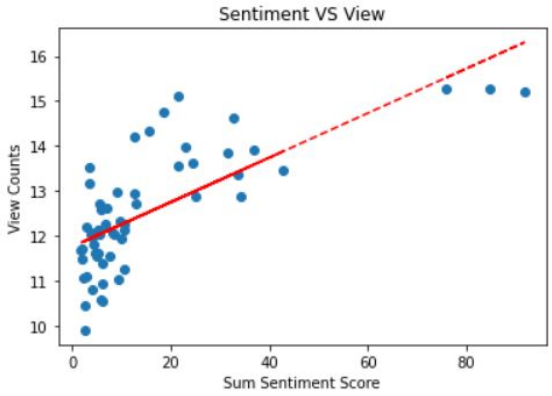


Figure 8: View Counts and Sentiment Score

From figure 7, we can see there is positive correlation between the view counts and M statistic, which validates our hypothesis. However, as for the correlation between sentiment score and view couts (figure 8), although there is a positive correlation shown in the graph, we should note that x-axis is the sum of the sentiment score of all comments instead of the average sentiment score. This is because the problem of unusual edits counts when the article is very popular. However, this is just to validate the hypothesis for view counts. When generating the final

measurement (The weighted sum formula), we still use the mean sentiment score because we try to make the sentiment score perform better in the low M region.

Weighted Sum Formula

After the above analysis, we generate our final weighted sum formula:

The level of controversiality = (Mean revert + Mean sentiment score)*log(View counts) + Log(M statistic)

Using this formula, we plot and compare the statistic line with the original measurement (M statistic), and find out **the most improvement is we successfully make the slope smoother** (see figure 9[3]). Although there is still a steep decrease within top 5% articles, we considered this more likely to be a feature in Wikipedia that most articles are relatively "boring" than a drawback of our measurement.

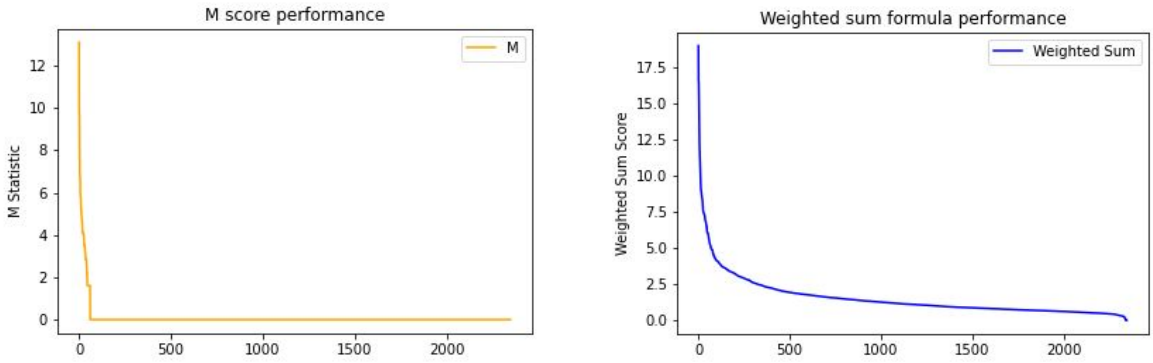


Figure 10,9: Comparison of old (left) and new (right) measurements

We can see a better version of measurement for controversiality has successfully been generated.

[3] Consider these two figures are one-dimensional, because the x-axis has no real meaning which just shows articles' ID in descending order by level of controversy.

Ours successfully detects controversy among “boring” articles while maintaining the M’s performance within popular articles.

Conclusion and Constraint

Overall our biggest success is to smooth the statistic line to make the controversy level for “boring” articles also visible and comparable. We first hypothesize and prove the validity of sentiment score, pageview and revert’s correlation with the real controversy level. Then we combine those factors to weigh the controversy accurately in both popular articles and relatively low controversial articles.

However, there are also constraints within those factors. First we use the “Vader“ sentiment model to measure the sentiment in each comment, but that model is not specifically built for the Wikipedia platform. Words such as “revert” ,”touch” are not detected as negative words as they should. Second, there is a small percentage of overlap between mean reverts and M, because M is calculated by the number of “mutual reverts”, which is a subset of all reverts. Yet we have not considered the influence of counting the intersection twice.

Future work

Looking through our project, you can get a new calculation method on defining the controversy of Wikipedia articles. After making those analyses on our weighted sum formula, it lets us know that there are still more things we can do in the future to let our work become better.

Firstly, we use limited data sources to finish our research. Specifically, the website called Wikimedia Data Archives contains all the edit history of Wikipedia. However, because of our limited hardware ability and the efficiency of our codes, we can only use 5 files in such large size. This may reduce the accuracy of our weighted sum formula. So, in the future, we want to get all of the edits with comments dump files from Wikimedia Data Archives to improve our analysis. With such support of having all edits with comments dump files, it can let us have more articles and more data related to those articles to make analysis. For example, our sentiment score for each article will be more accurate because each article will have all the comments they should have. So by doing this improvement in the future, it will let our weighted sum formula become more accurate for the articles in English Wikipedia area.

Secondly, besides just looking at the English Wikipedia articles, we can let our weighted sum formula become more adaptive for other Wikipedia articles. In other words, for each Wikipedia source, we can generate a weighted sum formula for that. And finally, we can have multiple formulas for detecting the controversy of articles in different sources. This may help us when we want to analyze other Wikipedia sources.

Third, there are also many possible EDAs (Exploratory Data Analysis) we should do to fully test the performance of our new formula. For example, we could split the score into a vector and aggregate articles by each factor to see each factor's performance and compare the difference in each article group. Such EDAs are going to help us further understand and improve the formula in the future.

Acknowledgments

The work is a research directed by the UCSD data science department. It is directed by our mentor Aaron Fraenkel and teaching assistant Kengchi Chang. Special thanks to those other group members who also give us valuable feedback and comments during our work.

References

Greenstein, Shane, and Feng Zhu. “Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia.” MIS Quarterly, vol. 42, no. 3, 2018, pp. 945–959., doi:10.25300/misq/2018/14084.

Halfaker, Aaron, et al. “The Rise and Decline of an Open Collaboration System: How Wikipedia’s Reaction to Popularity Is Causing Its Decline.” American Behavioral Scientist, 2012, doi:10.1177/0002764212469365.

Kittur, Aniket, and Robert E. Kraut. “Harnessing the Wisdom of Crowds in Wikipedia.” Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work - CSCW '08, 2008, doi:10.1145/1460563.1460572.

Sumi, Róbert, et al. “Edit Wars .” Privacy, Security, Risk and Trust , 2011, doi:10.1109/PASSAT/SocialCom.2011.47.