Leipzig University
Institute of Computer Science
Degree Programme Computer Science, B.Sc.

A New Controlled Dataset for Investigating Deliberation on Wikipedia

Bachelor's Thesis

Erik Jonathan Schmidt
Born Nov. 10, 1997 in Kulmbach

Matriculation Number 3719028

1. Referee: Prof. Dr. Martin Potthast

Submission date: March 30, 2022

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, March 30, 2022

.....
Erik Jonathan Schmidt

Acknowledgements

I would like to thank my supervisor Prof. Khalid Al Khatib and my referee Prof. Dr. Martin Potthast for their guidance throughout this project. Computational experiments in this work were done (in part) using resources of the Leipzig University Computing Centre.

Abstract

This thesis introduces a new corpus that is useful for distinguishing successful from failed deliberations on Wikipedia. Leveraging a resolving mechanism used in Wikipedia, we collect pairs of successful and failed discussions, both addressing the same discussion subject. This is essential for having a *controlled setting* and ruling out topical bias, allowing for more expressive analysis of how successful and failed deliberations differ. To collect these pairs, we develop an approach that utilizes the *Requests for comment* process within Wikipedia. This approach outputs 421 pairs that the new corpus consists of. An analysis of the collected pairs sheds light on the important role the first comment plays in the success of Wikipedia deliberations.

Contents

Chapter 1

Introduction

1.1 Motivation

Discussions are key means of human interaction as they build an interface to exchange different ideas and opinions, to understand the views and motives of others and to learn from each other.

Whether a discussion turned out successful is hard to tell in general. Success needs to be assessed differently for each kind of discussion. For example, in persuasive discussions, where the goal is to persuade the other party with a certain stance towards a particular topic, success could be defined by whether the other party adopts the stance or not.

In this thesis, the focus lies on *deliberative discussions*, where participants have the shared goal of making decisions. This goal is usually reached by bringing forth their ideas and by reviewing conceptions of others. Deliberating discussion is especially mandatory in groups, where many individuals need to make decisions together, that all can agree on. This applies to deliberative discussions extracted from Wikipedia talk pages, where a group of editors has the common goal of enhancing the quality and correctness of an article at hand. These discussions take place in written and anonymous form and therefore are accompanied by several shortcomings. Finding consensus on what the article is missing or how it should be altered as fast as possible is desirable in order to use the limited resources (time of editors) effectively. Deliberations that ended in consensus are viewed as successful, deliberations where the participants could not reach any agreement as failed.

The long-term goal is to investigate the deliberation strategies that lead to successful discussions. Each point made in a discussion can be considered to be one argumentative move and the sequence of these moves can be viewed as the *argumentation strategy* of the deliberation. We hypothesize that some argumentation strategies are beneficial for reaching consensus and other strate-

gies are prone to lead to disagreement. Once identified, these argumentative strategies could be leveraged to support ongoing deliberations. If failure-prone strategies are found, than following these argumentative paths that are likely to lead to failure can be avoided. On the other hand steps that are known to be advantageous can be proposed based on argumentative strategies that were found to be beneficial.

1.2 Approach

To find these inclined strategies the first step is to compile a dataset consisting of Wikipedia talk page discussions labeled as successful or failed. The discussions labeled as successful are expected to have reached consensus towards the end and the failed discussions end without agreement or without settling on a solution to the problem at hand. To draw reliable inferences about patterns of successful deliberation from this dataset it is necessary to rule out topical bias between the positive and negative subsets. Otherwise, revealed common patterns might not originate from the deliberations being successful, but from bias towards a particular subject that is overly represented in the set of successful discussions. Therefore we aim to only add pairs of a successful and a failed deliberation that tackle the same subject to the set. That way the distribution of deliberation subjects is the same in the successful and in the failed compiled deliberations. Controlling the variable of deliberation subject allows to create a *controlled dataset* to enable more informative deductions.

The (second step), that is not addressed in this work, is to supplement the discussions of this corpus with their argumentative strategies. That is labeling every successive utterance of the discussion with the strategic move it embodies. After these two steps, we can search for common successful and common failed argumentative strategies within the constructed corpus.

So the short-term goal addressed in the following is tackling the first step outlined above. To create a controlled dataset of successful and failed deliberations, 421 pairs of deliberations are collected by utilizing the context of discussions that are part of the predefined *Request for Comments* process within Wikipedia. How this process provides pairs of successful and failed deliberations with similar discussion subject is described in chapter 2. These pairs are presented in the new RfC-predecessor corpus and used for a brief analysis on differences between successful and failed deliberations. The conducted analysis (in chapter 4) illustrates the importance of stating the discussion issue clearly and comprehensibly in the first opening comment in order to deliberate successfully.

1.3 Structure

To illustrate the background of this work we show how deliberation is carried out on Wikipedia and how it is structured in the next chapter 2. Furthermore former research on on Wikipedia discussions is presented there. After stating the background we discuss how the new dataset can be compiled and depict the individual steps that were conducted achieve to the finished corpus in chapter 3. In chapter 4 an analysis on differences between successful and failed Wikipedia deliberations is performed based on the new collection of deliberations. Eventually chapter 5 briefly summarizes the work and highlights our key findings.

Chapter 2

Background and Related Work

2.1 Background

This chapter briefly introduces the Wikipedia talk pages and how editors can deliberate there. In the following related work about these deliberations is presented.

Deliberation is a form of discussion where participants work toward a shared goal. In order to reach this shared goal participants need to deliberate on how the goal is achieved best and subsequently act according to the found compromise. Contrary to that, there are discussions where agreement is not mandatory. When discussing opinions for example participants do not depend on agreement, here each participant has his own goal (defending his belief) rather than a shared goal (finding a solution to some issue that concerns all participants). Deliberation forms the base for any decision making in egalitarian organized groups. With a decision to make or a question to answer every peer of the group can state her view on the topic and can question the suggestions of others. This kind of discussion is referred to as deliberation and it needs to be consensus seeking, because in the end the decision needs to be backed by all participants. If such a deliberation converges from many individual standpoints towards one that all can agree on, then it can be seen as successful. If such consensus is not reached, then the group cannot present one standpoint as the standpoint of the group as a whole and the deliberation therefore failed.

Wikipedia is an online encyclopedia that consists of articles about various subjects, that are accessible in a 'wiki'-like manner and compile knowledge about a given subject from other sources. The writing and editing of theses articles is carried out by a community of editors. The English Wikipedia alone

has over 6 million articles[1] addressing subjects reaching from "Art and culture" to "Technology and applied science"[2]. Generally the term 'wiki' is used for a website that allows users to add, delete and edit contents[3]. This applies to Wikipedia where articles are written and edited by anonymous editors and can constantly be changed. Every article is therefore supplemented by a version history that keeps track of all these changes. In contrast to original research Wikipedia editors are supposed to follow the *No original research* policy. *No original research* means that editors need to back every fact that they add to an article with reliable secondary sources and must not include their own ideas or conclusions. Likewise statements that appear to be *original research* commonly get removed by editors when encountered.

By July 2021 the English Wikipedia counts about 388 thousand editors of which around 35.000 to 40.000 are active each month[4]. These editors are constrained to follow the Wikipedia guidelines that consist of rules, essays and defined processes for handling common situations like merging two similar articles into one or resolving a stuck discussion. These guidelines are meant to ensure the quality and correctness of the articles as well as instructing editors to communicate in a consensus seeking and professional manner. Most articles are created and extended step by step and by several editors.

Talk pages are in place for editors to deliberate about changes they made, suggest content that could be included or challenge the reliability of cited sources. These talk pages are accessible from the corresponding article and are also referred to as discussion pages.

Figure 2.1: The tab leading to the corresponding talk page of the Natural language processing Wikipedia article.

To start a discussion on a talk page one directly edits the source code of the page which is formatted using the Wikipedia specific "wiki markup". To

[1] https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons#Footnote_on_Wikipedia_statistics

[2] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Lists_of_topics

[3] https://dictionary.cambridge.org/dictionary/english/wiki

[4] https://en.wikipedia.org/wiki/Wikipedia:Wikipedians

organize the text added to a talk page into distinct discussions editors are encouraged to use a markup headline to start the discussion and indentation to arrange each following contribution: A new contribution is indented one level deeper than the one it is replying to. Editors sign their contribution with their editor name or IP address (see figure 2.2). This way discussions are readable

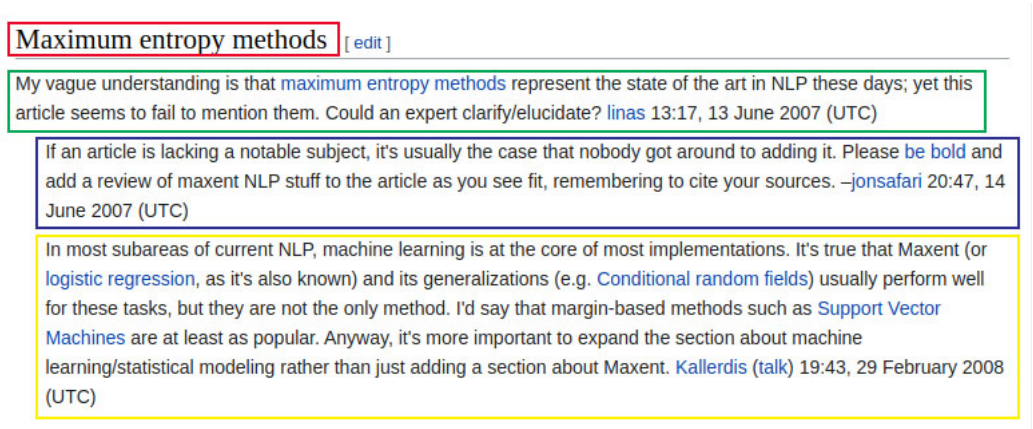


Figure 2.2: A discussion from a talk page. The headline in red, utterance one in green. First answer to utterance one in blue, answer two in yellow.

from top to bottom. This discussion structure is just a recommendation and not compulsory. Editors are able to edit old contributions even of other editors, although this makes understanding the pathway of the discussion much harder. Therefore discussions on Wikipedia talk pages are different in nature from discussions in forums or chats, where the chronology of contributions is imperative and contributions cannot be edited once posted.

Wikipedia talk pages have the single purpose to communicate about how to enhance the corresponding article. Therefore one crucial feature of discussions found on these pages is that the participants share the goal to improve the article, albeit their conceptions of what is an enhancement might differ. Furthermore editors are equal in that everyone can edit the article. If an editor wants his change to stay he needs to back his edit with argumentation and find agreement with the other editors. If he encounters editors with other perceptions then they need to settle on a compromise. Thus discussions on Wikipedia talk pages are viewed as consensus seeking argumentative deliberation. This perception is why deliberations that end in consensus are viewed as successful or failed vice versa.

Wikipedia records huge monthly traffic of editing. Each month since 2007 more than 4000 users made over 100 edits to articles and talk pages[5]. Among

[5] `https://en.wikipedia.org/wiki/Wikipedia:Wikipedians`

all those discussions there are many common topics like: "Should two articles on similar topics be merged?", that come up frequently. To help with these common issues guidelines and processes were developed, that define how to handle the problem at hand.

Requests for comment is one of these defined processes: "Requests for comment (RfC) is a process for requesting outside input concerning disputes, policies, guidelines or article content."[6]. Whenever an editor feels she is stuck in a discussion with others she can summarize the arguments brought forth and create a new discussion that starts with that summary on the same talk page. By marking this new discussion with a specific template the discussion appears on a global dashboard listing all ongoing Requests for comment. Editors with

```
=== RfC on capitalization of buildings ===
<!-- [[User:DoNotArchiveUntil]] 04:01, 19 February 2022 (UTC) -->{{User:ClueBot
III/DoNotArchiveUntil|1645243275}}
{{Rfc|style|hist|sci|rfcid=277D910}}
```

Figure 2.3: The RfC template underneath the headline of a RfC discussion.

no stake in the previous stuck discussion can now enter the deliberation and present a third opinion. With help of the attracted neutral editors the dispute described in the summary can be addressed and might get resolved. Talk page discussions including the RfC template are refered to as RfCs in the follwing.

The case study of the RfC process in [Im et al., 2018] found that 57.65% of RfCs end up getting formally closed by the addition of a summary statement resolving the dispute. Accordingly at least 57.65% of RfCs end in consensus and therefore are successful deliberations. Discussions that yield a RFC on the other hand have failed to reach consensus, because the editor who created the following RfC did not agree on the outcome and sought outside input. Thus the RfC process can provide pairs of a successful and a failed deliberation on the same subject. In Wikipedia there is no direct link implemented between RfC and the predecessor discussion. In this work an approach for linking these pairs is presented (See chapter 3.2).

[6] `https://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment`

2.2 Related Work

This section presents previous work that examined the Wikipedia RfC process, the success of discussions on Wikipedia Talk pages and datasets built based on the Wikipedia talk pages.

In the paper [Im et al., 2018] deeper insight into the progression of RfC discussions and the editors of such discussions are elaborated. To identify what hinders participants in resolving RfC discussions both a qualitative and a quantitative analysis are conducted. The foundation for this research were 10 interviews with editors that frequently close RfC discussions and a collection of 7,316 RFCs from the English Wikipedia. They found that 57.65% of the examined RfCs got closed formally by an editor who decided that this discussion succeeded. The remaining RfCs were either closed automatically after 30 days without activity or were withdrawn from the RfC processes, meaning that an editor removed the template that marked the discussion as RfC. The main focus of the paper is predicting whether a RfC discussion gets resolved by an editor or if it goes stale. A classifier for this prediction task was trained using features based on the discussion content as well as data about the editors. Feature analysis showed that the features of category *Participant Experience* are most predictive of the RfC going stale or not. In reference to the goal of compiling a corpus of successful and failed deliberations the formally resolved RfCs can serve as source for successful deliberations. Whether the ones that went stale failed to reach consensus is not evident. Furthermore a dataset constructed from successful and failed RfCs would not meet the specified requirement of controlling the topic or subject discussed about.

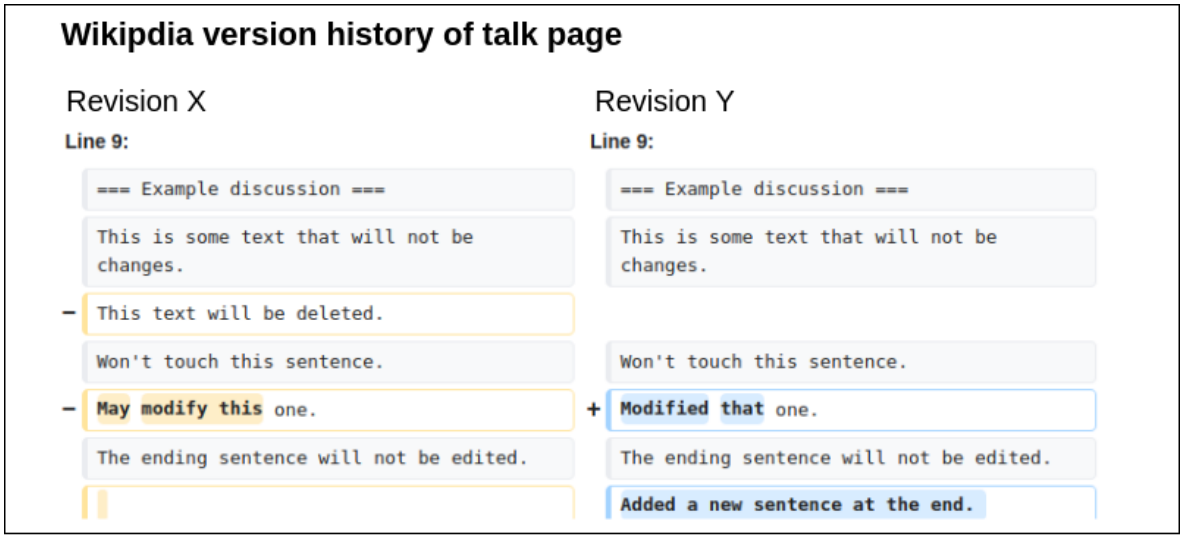
An other perspective on the success of discussions is obtained when analysing how successful a conversation was for the individual editors. In [Maki et al., 2017] the success of each editor is derived from the lasting changes he made throughout the discussion. Therefore the editor with the biggest impact on the article discussed about is said to be the most successful. For measuring the lasting changes an *editor success score* is introduced. Whether changes stay because editors achieved agreement or if a part of the participants is dissatisfied with the outcome is not captured. It is possible that the deliberation ended in dispute and the editor with lasting changes did not convince the others, but just was the most opinionated and resistant. Thus the success metric is not suitable to determine whether a deliberation was successful for the group of participants in a whole. Therefore it is not suitable to create the wanted dataset of successful and failed deliberations.

The topic of failed Wikipedia talk page discussions is touched in [Wang and Cardie, 2014]. They create a corpus of 3609 discussions that were tagged as dispute on Wikipedia and 3609 discussions that are not tagged as such. A

Support Vector Machine trained on this corpus predicts if a given discussion involves dispute or not with 80% accuracy. The prediction is based on sentence level sentiments as well as other features derived from the conversation text. The state of dispute is not considered in more detail for the discussions included in the corpus. That means a dispute discussion could end either with or without the dispute being resolved. Therefore the disputed discussions do not meet the assumptive definition that failed discussions end without agreement.

The paper [Al-Khatib et al., 2018b] focuses on a turn level analysis of discussions, therefore discussions are split into individual utterances brought fourth by the participants. About 6 million talk page discussions are extract from a Wikipedia dump and split into 20 million individual turns. Furthermore tags, shortcuts, templates and links that are used in a particular turn are covered when given. These extracted discussions are compiled to the *Webis-WikiDiscussions-18* corpus. A concept is presented that deduces the argumentative purpose of a turn from the tags, shortcuts, templates and links it contains. The concept suggests thirteen such argumentative moves, which are then assigned to 200 000 turns from the former dataset. The proposed concept for identifying argumentative moves can be used for a deeper analysis of successful and failed deliberations in future work.

In contrast to the datasets created from a single Wikipedia dump a complete corpus of talk page discussions is created in [Hua et al., 2018]. Deliberations that got deleted by the time of corpus construction get captured as well, by processing the version history of each talk page instead of using only its current markup code. This approach finds 91 million deliberations. While the core objects in the *Webis-WikiDiscussions-18* are the discussion turns the *WikiConv* corpus consists of edit actions. For a given deliberation all changes to the discussion text can be reviewed in the talk page version history. In the *WikiConv* corpus these changes are parsed into the edit actions: Addition, deletion and modification. Addition means that between version X and the following version Y new text was added while all text from version X is still present in Y. Deletion on the other hand means that X contained some text that is missing in Y and Y contains no text that was not already there in X. Modification refers to replacing the exact text of a former addition. It is possible that changes between two successive version are parsed into several of these edit actions (See Figure 2.4). All actions that resulted from a change get attributed to the editor who performed the change.



ID	Type	Content	Editor
1	Deletion	"This text will be deleted."	User A
2	Modification	"May modify this" -> "Modified that."	User A
3	Addition	"Added a new sentence at the end."	User A

Figure 2.4: Changes between two successive versions X and Y of a example discussion as viewed in the Wikipedia version history tab. Bellow a schematic depiction of the edit actions that capture the changes from version X to Y.

Chapter 3

Corpus Construction

In this chapter the process of creating the RfC-predecessor corpus is explained. In the first section different reasonable approaches towards compiling a controlled dataset of successful and failed deliberations are explained. The second section depicts the creation of a corpus with deliberations taken from the RfC process.

3.1 Utilizing the RFC Process

[Al-Khatib et al., 2018a] introduces two publicly available datesets composed of talk page discussions. The *Webis-WikiDiscussions-18* corpus consists of 5 941 534 discussions extracted from talk pages in the english Wikipedia dump from March 1st, 2017. The "wiki markup" files of single talk pages were disassembled to discussions with the use of regular expressions for finding start and end of the discussions. In a subsequent step the *Webis-WikiDebate-18* corpus was created. It extends the former corpus by assigning labels to around 200 000 of the discussion moves. Discussion move here refers to a single individual utterance made by an editor in the course of the discussion, a discussion can therefore by represented as a sequence of moves. The labels reveal what function an individual move has for the overall discussion. Meta data, such as Wikipedia specific shortcuts and links, was inspected for each move to find the argumentative label that fits. If an editor added a Wikipedia shortcut to guidelines about validity of sources to his utterance, than this accounts for assigning the label *"Verifiability and factual accuracy"* to this move. One of these labels is *"Finalizing the discussion"* and it was assigned to 622 utterances that include meta information about reporting, committing or archiving a discussion. Committing and archiving a discussion could be interpreted as sign of successful closure, reporting on the other hand could mean quite the opposite as well. This said, the corpus was created from a Wikipedia dump

and thus captures discussions that were displayed on talk pages at time of dump creation only. Discussions that took place earlier and were archived or deleted before corpus construction are not considered. To use all resources available the whole version history of each talk pages needs to be examined to extract all discussions.

As the labels from *Webis-WikiDebate-18* appear to be unsuitable for identifying successful and failed deliberations an other approach is developed. Therefore meta data that provides context to certain Wikipedia talk page discussions is evaluated. The before mentioned RfC process embeds discussion in such context: A discussion posted to the RfC notice board is from the outset known to address an issue that could not be solved in a former discussion. As mentioned above, former research showed that a large share of deliberations posted to the RfC notice board conclude with a formal resolution, which matches our definition of a successful deliberation as a deliberation that ended in consensus. Furthermore the context provided by the RfC process can be leveraged to find failed discussions as well. The RfC process is installed to resolve issues that were discussed in a first deliberation without the editors coming to agreement. In such a case one of the editors opens a second deliberation that states the issue they failed to resolve in the first place and invites other impartial editors to help. Consequently the first deliberation did not reach consensus and therefore failed, according to our definition of failure. As displayed the RfC process offers a point of reference for finding both successful and failed deliberation.

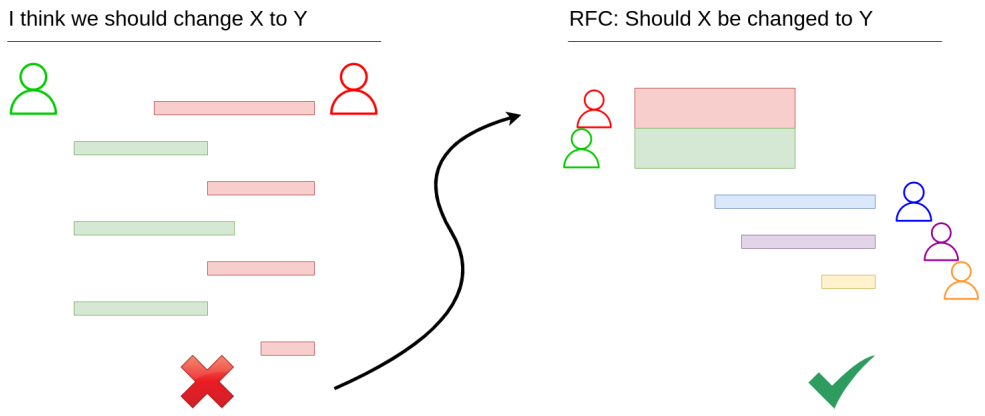


Figure 3.1: Schematic depiction of a failed predecessor discussion and the RfC discussion subsequently posted to the RfC notice board

A further advantage of the approach to investigate the RfC process for finding successful and failed deliberations is that we can find pairs, one successful and one failed discussion, where both address the same issue and both take place within the same domain. Same domain means they both discuss

the same article that the talk page corresponds to and more even they address the same issue that could not be resolved in the preceding discussion and is sought to be resolved in the RfC deliberation (See figure 3.2).

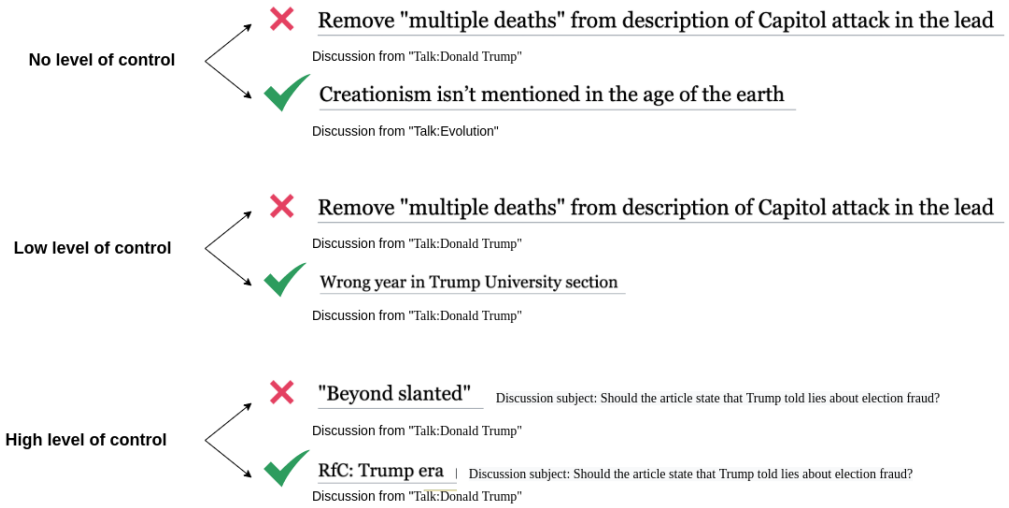


Figure 3.2: Illustration of three deliberation pairs. Each pair consists of one successful and one failed deliberation. In the first pair the deliberations are take from unrelated talk pages, therefore the discussion subject differs considerably. Both deliberations of the second pair stem from the same talk page, therefore the discussed content is more similar. The third pair exhibits high level of control regarding the discussion subject: Both deliberations discuss the exact same issue.

Assembling the corpus from such pairs would fulfill the requirement to create a controlled dataset, in that there is no bias towards a particular topic or subject within the successful deliberations that is not given among the failed deliberations as well. When derived from this corpus, conclusions about how successful deliberations differ from failed ones would not run the risk to stem from different topical bias in the two subsets. It must be said however, that the deliberations in the successful subset are different from the failed ones in nature: The successful deliberation, being tagged as RfC, started with a consciousness that there is an issue that caused disagreement previously and that the goal is to resolve this particular problem. In the predecessor deliberation this consciousness of a sever disagreement that has to be overcome is missing. This difference in the frame of the conversations might blur the expressiveness of conclusions drawn from the corpus and hinder transfer to deliberation in general.

To keep talk pages manageable editors often archive or delete resolved deliberations. Therefore it is mandatory to investigate the whole version history of a talk page when searching the predecessor of a RfC, as the predecessor

might got removed at some point. Using datasets created from a particular Wikipedia dump, such as *Webis-WikiDiscussions-18* is expected to lead to less predecessor-RfC pairs than revising the version history of the corresponding talk page.

To sum up: The acquired approach for corpus creation is to collect all RfC deliberations of the English Wikipedia as successful discussions and to pair each RfC with its preceding failed discussion to compile the controlled dataset of successful and failed deliberations.

3.2 Collecting RfC Deliberations

As explained before, the individual utterances of editors within a talk page discussion are not posted to a chronologically ordered thread or chat, but are added to the talk page markup file through out edits. It is recommended to add a new contribution below the last previous utterance, but that is not compulsory. For example an editor A could delete parts of the preceding utterance by editor B or add his own utterance just above the one by editor B. The chronological progression of a conversation can only be reconstructed with certainty by checking the version history of the talk page. The version history comprises every change to the talk page with timestamp and editor. The differences between two versions of the same article can be viewed in a Git like manner. It is common practice that all the text that makes up a deliberation is deleted by one of the participating editors once settled on a solution, so that the talk page does not get polluted with outdated and resolved conversations. This is why a Wikipedia dump, that saves the state of Wikipedia at one point in time, does not capture all discussions that ever happened.

The WikiConv corpus contains all English talk pages with exhaustive version history until July 2018. The corpus identifies the individual conversations and the changes to conversation text. Differences of a conversation between version X and the following version Y of the talk page get parsed into one of five edit actions (See figure 2.4) marked with an identifier that is unique for that particular conversation. All five edit action objects have the same format displayed in figure 3.4 and therefore can be stored in one large table of 220 million rows representing the edit actions. That way corpus memory space is reduced dramatically compared to storing every single version of each talk page. By performing each edit action of a conversation in chronological order the conversation text can be reconstructed for every point in time. In total WikiConv contains 90 930 244 separate conversations and therefore is a richer source of deliberations than the Webis-WikiDiscussion-18 corpus with its 5 941 534 discussions.

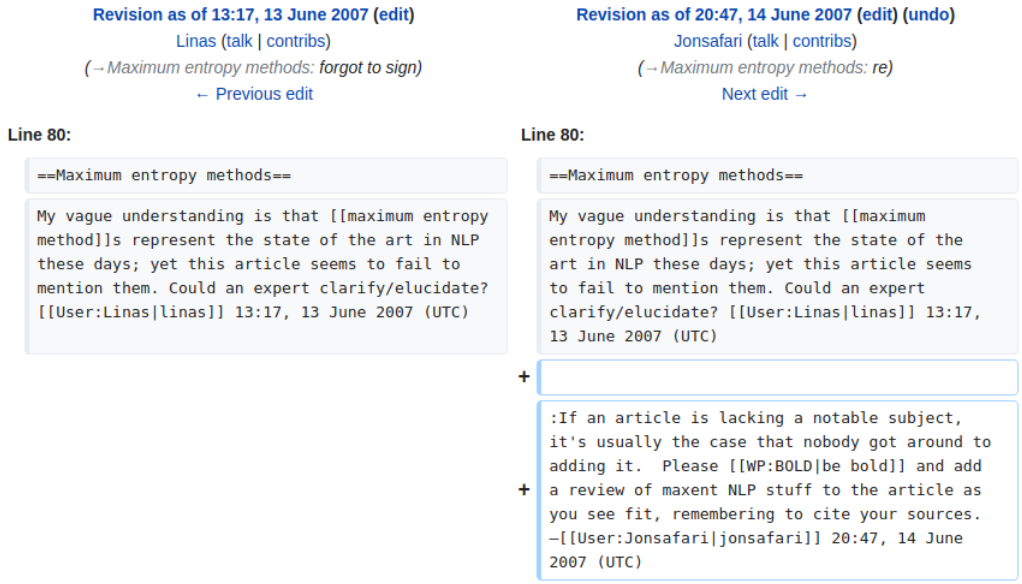


Figure 3.3: Comparison of two subsequent versions of the Natural Language Processing article talk page. The change is part of the discussion in figure 2.2. As recommended editor "jonsafari" adds his answer to comment by "linas" by inserting text below the last utterance to create an understandable conversation.

id	conversation_id	page_title	indentation	replyTo_id	content	cleaned_content	user_text	rev_id	type	user_id	page_id	timestamp	parent_id	ancestor_id

Figure 3.4: Scheme of the *WikiConv* corpus table scheme

The WikiConv corpus is publicly available in Cornell ConvoKit format and on Google Cloud Storage as single table. The table on the Google Cloud Storage can be accessed through the Google Cloud API with the identifier 'wikidetox-wikiconv-public-dataset' and is split into 500 distinct lists. All 500 tables together the dataset is 750 GB in size.

Given this dataset the aim is to extract all edit actions of deliberations that contain the RfC template (See figure 2.3) and the actions of discussions that could be the predecessors of these. The extraction is carried out as follows:

1. As first step each of the 500 edit action lists is downloaded one at a time. For each list all actions get scanned for the RfC template and the talk page ids of actions that contain the template are saved. After all actions got scanned the table gets discarded.

2. At next each table is downloaded again and with the before stored list of talk page ids all actions of these talk pages get extracted and stored. Then the original table is discarded again. Downloading each table twice

allows to reduce the memory needed for filtering at the cost of taking twice as long.

3. After this filtering process the actions of 8606 talk pages are left for further steps. The total table size is reduced to 37 GB. These 8606 talk pages contain 17 542 distinct deliberations with the RfC tag.

4. Given the conversation ids of these 17 thousand RfC deliberations the latest state of wiki markup is reconstructed for each by performing its edit actions in chronological order. For many this results in blank text due to final deletion edit actions. Sorting out all deliberations that are reconstructed to blank text with this approach decreases the number of RfCs to 11 554.

Implementing a solution to identify when a RfC reached its concluding state before being deleted or trimmed would be beneficial but was not part of this work.

3.3 Compiling RfC-Predecessor Pairs

Now all RfC discussions that exist on Wikipedia until July 2018 are obtained. The ones that could be reconstructed to a readable conversation are collected together with accompanying discussions from its talk pages, consequently the next step is to identify the predecessor deliberation of each RfC in order to get pairs of successful and failed deliberations. Wikipedia has no direct link installed between RfC and predecessor. A process for linking a RfC deliberation with an other deliberation from the same talk page that is thought to leaded to the RfC is depicted in the following.

3.3.1 Finding Predecessor Candidates

The talk pages that remain contain 447 deliberations on average. So, on average, for each RfC one of 446 (all deliberations of the talk page minus one as the RfC cannot be its own predecessor) needs to be picked as predecessor deliberation. As first step of this picking process conversations get separated out, that logically cannot be the predecessor.

1. First of conversations that were carried out after the RfC deliberation started are removed from the candidates.

2. Secondly deliberations that have no editor in common with the RfC are sorted out, because according to the defined RfC process one of the

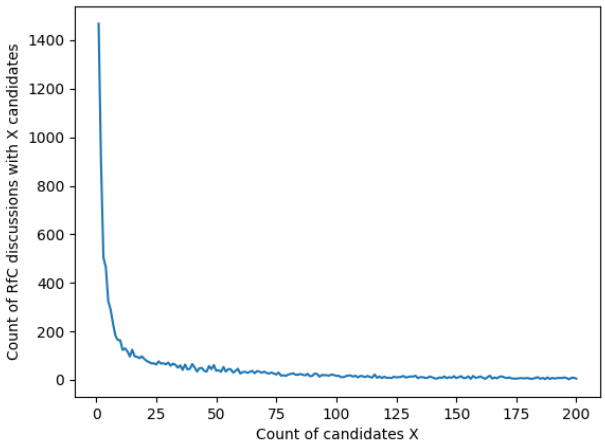


Figure 3.5: The distribution of candidates among the 11 554 RfC discussion reconstructed to none blank text (truncated after candidate count bigger 200)

participants who finds herself stuck in a discussion initiates the RfC discussion and therefore is present in both the predecessor and the RfC deliberation.

3. Thirdly a time limit of 7 days between end of the candidate and start of the RfC is introduced.

These three conditions reduce the average number of candidates per RfC to 159.

The number of candidates is rather small for most RfCs (See figure 3.5), but the average candidate count is pushed by a long tail of RfCs that have way more candidates than average. Examining the distribution of RfCs and their corresponding candidates reveals that 1468 RfCs have only one candidate for being its predecessor and 3333 RfCs have between one and four candidates.

These RfCs with up to four candidates are used for a first manual examination. From each set of RfCs with one, two, three or four candidates ten RfCs with corresponding candidates were randomly picked. For each of these 40 picked RfCs the reconstructed deliberation of the RfC and its candidates are read. Then, based on subjective judgment, either one of the candidates is picked as predecessor if it is recognisable that it has initiated the RfC discussion. Where there is not suiting precursor among the candidates no deliberation is picked as predecessor. Note that the candidate conversations are reconstructed to the latest state before the RfC began, as this is the state that is assumed to have led to the RfC. Table 3.1 shows the results of this manual assignment. For ten of the 40 inspected RfC none of its associated candidates

Table 3.1: Results of manually picking predecessors for 40 randomly picked RfC deliberations. Conducted on RfCs with one, two, three and four candidate discussions, ten of each group.

candidate count	predecessor found	predecessor not found
1	8	2
2	7	3
3	10	0
4	5	5

related to the RfC as predecessor. This large share of RfCs with no identifiable predecessor could originate from unsuitable reconstructed discussion texts, because the reconstruction of these texts is premised on a simplification. When determining whether a certain deliberation was the predecessor of a RfC at hand or not, the candidate text is reconstructed to its latest state right before the RfC started. So only the text displayed in this latest version is considered for the manual assignment; if a candidate was mostly altered or deleted from a talk page before the RfC discussion started, then all altered and deleted parts are not considered when assessing the candidate. This shows the difficulty that comes with talk pages discussions not having a compulsory chronological progression, but a loose structure captured in the version history. An other cause seems to be that the RfC process is not always used as intended and defined. Some discussions get tagged as RfC right away to get more attention from the start and therefore have no predecessor. These insights considerably hamper the process of linking RfCs with its predecessors. A first implemented linking mechanism that ranks all candidates for a given RfC and picks the best scoring as predecessor will fail for all RfC that miss the actual predecessor in the list of candidates. Based on the findings from the manual inspection, this would affect a quarter of all RfCs in therefore is not applicable.

3.3.2 Investigating Predecessor Candidates

Guided by these insides the task of picking one or none predecessor for a RfC is simplified by focusing on RfCs with just one candidate. With this narrowed focus the process of choosing one or none predecessor deliberation from many candidates is boiled down to deciding whether the only candidate is the predecessor or not. Starting with 1468 RfCs that are paired with one candidate after the initial filtering, at first those are sorted out where the candidate conversation is blank when reconstructed. As mentioned before the text of a candidate discussion is reconstructed by applying all edit actions of that discussion in chronological order up to the starting point of the RfC

discussion. That means that discussions that were reconstructed to blank text were not blank at some other point of time.

After sorting out empty discussions those are excluded where either the RfC or the candidate text consist of less than two utterances, as a conversation of just one utterance can hardly be called a deliberation and consequently has no use for investigating deliberation. After preprocessing the 1468 RfC with exactly one candidate 613 RfC-candidate pairs remain.

For a given pair of RfC and candidate different features, derived from either meta data or the conversation content, are assessed on their value for identifying whether a pair contains the RfC and its actual predecessor or not.

Meta Features

The considered meta features are:

- The **time** between the end of the candidate and the start of the RfC. The assumption is that the shorter the time between candidate and RfC the higher the chance that it is the actual predecessor
- The **participant overlap** calculate as Jaccard similarity of the participants of the RfC and the candidate. A bigger overlap in participants is expected to correlate with the candidate being the predecessor deliberation.

Content Features

The considered content features are:

- The **bag of words** similarity between candidate text and RfC calculated as Jaccard similarity.
- The **link overlap** given as Jaccard similarity between links that occur in the candidate and in the RfC.
- The **shortcut overlap** calculated as Jaccard similarity between Wikipedia shortcuts appearing in the candidate and the RfC.
- The **keyword overlap** calculated as Jaccard similarity between the top five keywords of the candidate and the RfC.
- The **average best matching sentence similarity** calculated from sentence embeddings . For each sentence of the RfC the sentence of the candidate with highest cosine similarity is determined. The average of these cosine sentence similarities is the average best matching sentence similarity.

- The **single best matching sentence similarity**. For each sentence of the RfC the sentence of the candidate with highest cosine similarity is determined. The highest similarity found is the single best matching sentence similarity.

Feature Analysis

After computing these similarity features for each RfC-candidate pair the obtained values were normalized to take values between zero and one so that they are comparable to each other. The hypothesis that higher similarities are found among the actual predecessor pairs is verified by comparing the average values (See table 3.2).

It is analysed for which similarity feature the average value of actual RfC-predecessor pairs differs the most from the average of pairs without predecessor. This reveals that the single best matching sentence similarity is the best indicator for detecting predecessor discussions within the given RfC-candidate pairs. This feature is then used to extract a subset of pairs that most likely consist of a RfC and its preceding discussion from the 613 RfC-candidate pairs. To do so the pairs are grouped into ten subsets by their sentence similarity score. For each subset ten random samples are picked and assessed. It shows that the share of RfC-predecessor pairs is high among the four subsets of pairs with the highest similarities.

3.3.3 Picking Predecessors

The four subsets with highest similarities are combined to generate the RfC-Predecessor corpus. Concluding from the assessed pairs this combination of these top four subsets is estimated to contain $34/40 = 85\%$ actual RfC-predecessor pairs. The combination results in a compiled set of 421 samples.

Table 3.2: Comparing the average of normalized feature values for pairs containing the RfC predecessor(positive) and pairs that do not(negative).

similarity feature	average of positives	average of negatives	difference
time	0.117	0.059	0.058
participant overlap	0.215	0.200	0.015
bag of words	0.224	0.148	0.076
link overlap	0.012	0.0	0.012
shortcut overlap	0.0	0.0	0.0
keyword overlap	0.117	0.059	0.058
average sentence match	0.124	0.037	0.086
single sentence match	0.138	0.047	0.091

This RfC-predecessor dataset is used to train different classifiers in the following with the goal to extend the dataset further.

Predecessor Classifiers

The classification task is to detect whether the candidate deliberation is the predecessor of the RfC for a given RfC discussion and one of its predecessor candidates. Training such a classifier aims at extending the given RfC-predecessor corpus by identifying the predecessor for other RfCs from the 11 554 RfCs collected before.

The RfC-predecessor corpus contains pairs: A RfC and the deliberation that (most likely) preceded it. These pairs are referred to as positive samples for the training process in the following. To obtain the training set one negative sample is constructed for each positive one. The positive samples are created from RfCs that are associated with only one predecessor candidate conversation, consequently there is no other discussion one the same talk page that qualified as predecessor candidate in the first place. So for each positive sample all discussions from its origin talk page are collected, besides the RfC itself and the predecessor discussion contained in the positive sample. One of these deliberations is picked and paired with the RfC to create the corresponding negative sample: The RfC and a deliberation that (most likely) has not preceded it.

To decide which of the discussions to pick each of the so collected is compared to the actual predecessor deliberation of the corresponding RfC. It is assumed that using deliberations that are more similar to the actual predecessor in form and content will lead to a better performing classifier. Therefore the collected discussions are reduced to five at most by choosing the ones that have similar length as the predecessor. Secondly a single discussion is picked that has the highest similarity in content. The similarity in content is calculated as the average of the top five cosine similarities between sentence embeddings of the predecessor and the pertained discussion. Following this procedure negative samples can be created for 400 of the 421 positive samples. For 21 of the positives there is no other none blank discussion found on the talk page than the RfC and its predecessor.

A manual evaluation is conducted for 10 of the 400 RfCs that are used in both one positive and one negative sample. It shows that four of the ten selected not predecessor discussions can not be interpreted es such with certainty. Three of these problematic discussions seem to be part of the RfC itself and were wrongly parsed into separate discussions in the WikiConv corpus. The other one is very similar to the actual predecessor of the RfC. Editors might have disputed their concern in those two discussion before opening the RfC.

To overcome the issue of ill-parsed RfC discussions a new condition is introduced for picking a discussion for a negative sample: The discussion needs to have ended before the RfC has started. This reduces the RfCs usable for one positive and one negative sample further to 319. The new condition eliminates the problem of ill-parsed RfC discussions and in a second evaluation of ten randomly picked RfCs and their paired discussions only one is paired with a predecessor that seems inadequate. This leaves a training set of 319 positive and 319 negative samples that is later used to train several classifiers.

In order to find which classifier performs best on the task a test set is composed from 30 manually inspected RfCs and their candidates. Ten RfCs are picked from the set with exactly two candidates where the predecessor can be indentified with high confidence. The RfC and the predecessor discussion make up the positive test sample, the negative is constructed from the RfC and the other candidate discussion. In the same manner ten RfCs with three candidates are used to construct ten positive samples and one of the remaining two discussions is picked randomly to create the negative. The same procedure is applied to ten RfCs with four candidates. That way a test set of 30 positive samples and 30 negatives is obtained.

Given the training set from above and the explained test set different models are trained on the classification task.

BERT The first model is the transformer based BERT model presented in devlin2018bert. Transformer models achieve state of the art results on several natural language processing(NLP) tasks, such as automated question answering and text classification. A core component of BERT is its self attention mechanism which enables it to capture contextual relations between tokens of a sequence or sentence. Pretrained BERT models are available in many variants and can be fine-tuned for custom tasks. To obtain the first transformer based classifier the pretrained *BertForSequenceClassification* model available in the *Huggingface* library is fine-tuned with the training set. BERT limits the sequence length to 512 tokens because training costs grow quadratically with sequence length due to [Beltagy et al., 2020]. Therefore the training and test samples that exceed the maximum sequence length are trimmed. Candidate discussions get reduced to the last 512 tokens of the discussion text. The RfC discussions are cut after the first 512 tokens, so that the BERT model is fed with the ending sentences of the candidate discussion and the start of the RfC. After training BERT with the trimmed training set it is tested on the trimmed test set. It achieves an accuracy of 50% by assigning the same label to all samples, meaning it does not beat the baseline of randomly assigning one of the two possible labels to the samples. Evaluating the learning curve shows, that both training loss and validation loss decrease in very small increments.

Therefore it is assumed that the training set of 638 instances is to small to train the model and that curtailing the discussion texts hinders the training.

LONGFORMER To suit the BERT model the samples were trimmed and therefore the model did process complete discussions. To overcome this issue a second model is trained that can process longer sequences. LONGFORMER presented in [Beltagy et al., 2020] is a variation of the BERT model that limits the self attention mechanism to a sliding window of fixed size. That way the training costs grow linear with sequence length instead of quadratically as with BERT. *Huggingface* provides the pretrained *LongformerForSequenceClassification* model that is fine-tuned with the complete texts of deliberations in the training set. After fine-tuning the pretrained model for the classification task it is evaluated on the uncurtailed test set. It achieves the baseline accuracy of 50% by assigning the same label to all samples. Eventually both transformer based models were not able to learn the classification task on the given training set.

Logistic regression While transformer based models are directly trained on text learning algorithms like logistic regression and support vector machines rely on numeric features that fit the task. Therefore the eight features inspected earlier are calculated for all samples in the training set and in the test set. Subsequently these feature vectors that represent the original training samples are split into a training subset of 80% and a validation subset of 20%. The *sklearn*[1] library is used to fit the parameters of the logistic regression model to the feature vectors in the training subset. After fitting the parameters the hyperparamter C which determines the level of regularization is picked according to its performance on the validation subset. The best performing model trained as described achieves an accuracy of 57% on the feature vectors representing the test set samples.

Support Vector Classifier The training subset and validation subset used for training the logistic regression are also used to train the *Support Vector Classifier* provided by *sklearn*[2]. First the training subset is used to to fit the parameters of the support vector machine. In a second step the validation subset is used to pick the best performing value for the regularization hyperparamter C. Similar to the logistic regression model the the SVC achieves an accuracy of 57% on the test set. Besides the small size of the training set

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`

another explanation for the poor performance of logistic regression and SVC is the difference between the assembled training samples and the test samples. In the training set the positive samples are created from RfCs and the only discussion that passed the preconditions for being the predecessor. The negative samples on the other hand are constructed with discussions that did not pass these preconditions. In contrast to that in the test set both positive and negative samples are constructed with discussions that fulfill the precondition for being the actual predecessor. It is thinkable that this leads to easier distinguishable positive and negative samples in the training set than in the test set, so that learning on the training set does not translate well to the samples in the test set.

Concluding the training process it can be said that none of the obtained classification models is suitable for extending the dataset of RfC and predecessor discussions. The better performing models with 57% accuracy would not be precise enough to detect actual RfC predecessors that can be assumed to be such without further inspection. Given that the dataset cannot be amplified with the trained classifiers the already obtained data set of 421 RfC predecessor pairs is left as it is and inspected in the following. This concludes the construction of the RfC-predecessor dataset.

Chapter 4

Evaluation and discussion

In this chapter we analyse how the successful and failed deliberation in the obtained dataset differ and explain where these differences could originate from. Secondly we depict how future work towards the goal of understanding successful argumentation strategies could look like and discuss shortcomings of the presented corpus.

4.1 Evaluation

The depicted corpus construction process resulted in a set of 421 successful and 421 failed deliberations. As opposed to datasets from related work this new dataset rules out overrepresentation of particular topics in either the success or failure subset to allow for more reliable deduction on differences between successful and failed deliberation on Wikipedia. Given this dataset a first shallow analysis on differences between discussions in the two subsets is conducted. For each feature examined in this analysis the average among the successful deliberations is obtained as well as the average among the failed. The considered features are:

- The count of participants involved in the deliberations.
- The duration of deliberation in days. For each discussion its duration is given by the timestamps of its first and last corresponding edit actions. This includes late edit actions that removed discussions from the talk page and therefore in many cases is longer than the actual deliberating.
- The text length of the deliberation based on the reconstructed conversation as used in corpus construction.
- The count of subsequent comments as visible in the reconstructed discussion.

Table 4.1: Average values obtained from the successful and failed deliberations (Rounded to two decimals).

Feature	successful	failed	ratio
Participants	7.33	4.11	1.78
Duration	168.66	235.72	0.72
Text length	5617.90	4543.19	1.24
Comments	21.83	19.47	1.12
Changes	19.50	14.45	1.35
Edit actions	24.96	17.05	1.46
Average indentation	1.10	1.62	0.68
Max indentation	3.89	4.63	0.84

- The count of changes to the deliberation text, determined as count of distinct timestamps in the edit actions associated with deliberation.
- The count of edit actions in *WikiConv*.
- The average indentation depth of comments in a deliberation. This is again obtained from the reconstructed discussion as used in corpus construction.
- The level of indentation of the deepest indented comment in the deliberation.

Average values of these features in both the successful and the failed subset are illustrated in table 4.1. The analysis shows that more editors participate in the successful deliberations or in RfCs respectively. While the failed or predecessor discussions are visible only on one talk page the RfCs are listed on a global notice board with the goal to attract impartial editors to participate. Hence the difference in participating editors reflects the functional principle of the RfC process. Further the failed discussions continue for longer, compared to the successful RfCs. Similarly this is assumed to be an effect of RfCs being listed on the RfC noticeboard where they are less likely to be left stale before eventually deleting them because they are visible to more editors.

Text length, count of comments, count of changes and count of edit actions all are slightly higher for the successful deliberations. However, compared to the difference in participants the increase is rather low which means that in the successful deliberations each editor makes up for a smaller part of the conversation than in the failed.

Both the maximum and average level of indentation show to be higher in the failed discussions opposed to them being shorter and having fewer comments. When an editor brings forth a new comment in a discussion the level of

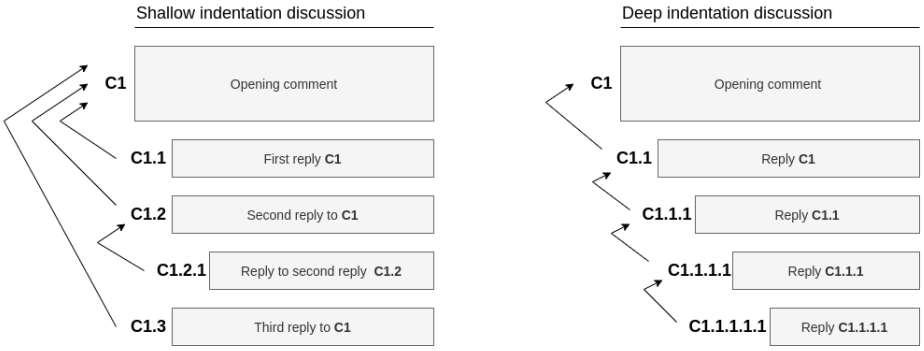


Figure 4.1: Schematic depiction of a Wikipedia talk page discussion with low level of indentation(left) and one with deep indentation(right).

indentation indicates which preceded utterance this new comment is responding to. Consequently deeper indented comments means longer sequences of directly referencing utterances. Shallow indentation is observed where more comments refer to the opening statement of the discussion (See figure 4.1). Therefore deeper indentation can be interpreted as editors getting sidetracked from the original matter. Moreover, deep indentation could be a result of the discussion issue not being sufficiently articulated in the opening comment of a deliberation, so that a back and forth of clarification is needed before editors can work on a solution. RfC discussions start with a short summary of the former issue that could not be resolved previously. Therefore the RfC opening comments can be assumed to be more comprehensive and clear than in the average talk page discussion, which would explain the lower level of indentation in the successful deliberations. These findings can be seen as a hint to the importance of opening comments that clearly point out the problem at hand or the question to answer in order to deliberate successfully. However certain conclusions require deeper and more focused analysis than the presented shallow sweeping blow.

4.2 Future Work

The newly constructed dataset aims to help indentify successful deliberation strategies. A next step towards that goal is to augment the collected deliberations with the argumentative move labels presented in [Al-Khatib et al., 2018a]. A classifier trained on the *Webis-WikiDebate-18* corpus is expected to be able to assign these labels to the presented dataset, so that common deliberation strategies, meaning sequences of these labels, can be detected among the successful and the failed deliberations. A qualitative analysis of deliberations

that follow these common strategies is assumed to provide insides into how effective and successful deliberation can be facilitated and which deliberative strategies better be avoided.

Furthermore extending the dataset with more controlled pairs of successful and failed deliberations is mandatory in order to achieve more representative results. Enhancing the presented feature based classifiers (logistic regression and SVM) would help to find more RfC predecessor pairs among the total 17 thousand RfCs available in the *WikiConv* corpus. To do so one could for example examine how the confidence scores relate to correct predictions in order to pick those pairs that the classifiers are quite certain about. In addition to that one could develop new features to help the classifiers.

Besides compiling more pairs from the RfC process other processes defined for common situations on Wikipedia talk pages can be leveraged. For example the *Dispute Resolution Board*(DRN) is installed to resolve stalemate discussions where editors disagree on article content in particular. Assumably pairs of successful and failed deliberations can be compiled from this and other processes by following a similar approach as depicted for the RfC process.

Supposing that the dataset can be extend one way or the other a distant goal is to build discussion assisting tools based on such an extended dataset. Given the argumentation strategies of thousands of successful and failed deliberations rather than hundreds one could derive frequent patterns of argumentative moves among the successful deliberations in order to reveal promising argumentation strategies. Aware of these successful patterns an assisting tool could observe ongoing discussion and suggest which argumentative move could get brought forth next in order to follow one of these successful deliberations strategies. That way a corpus of successful and failed deliberations paired with the before mentioned argumentative labels could be used to make Wikipedia talk page discussions more effective and goal oriented. Furthermore knowledge of successful argumentation strategies in Wikipedia discussion might carry over to online deliberation on different domains and deliberation in real life conversation.

4.3 Discussion

When using the dataset for an investigation of patterns in successful and failed deliberation it is important to keep in mind the different nature of the failed deliberations compared to the successful ones. As explained before editors enter a RfC deliberation that is listed on the global RfC noticeboard aware that they need to find a tradeoff solution for an issue that previously caused some sort of trouble. Therefore the context of these deliberations is different

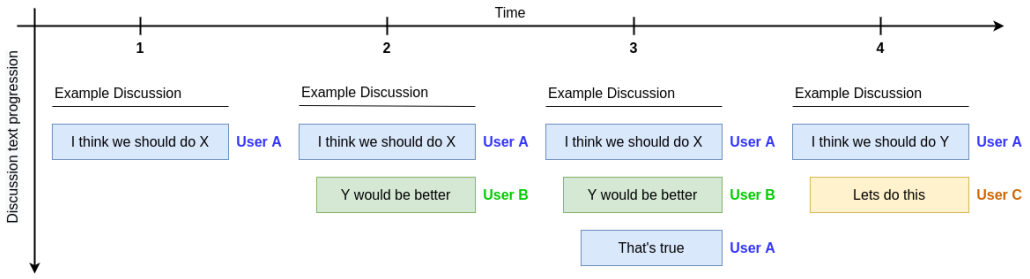


Figure 4.2: Schematic depiction of the progression of a Wikipedia talk page discussion: (1) User A opens the discussion. (2) User B replies. (3) User A agrees. (4) User A adjusts the opening comment with respect to the suggestion of User B. User A deletes deprecated utterances. User C replies to altered opening comment.

from the context the failed predecessor deliberations happen in: The failed predecessor discussions take place only on a particular article talk page without any frame or context that could inform editors of an upcoming dispute. So, in contrast to the discussion subject, the context is a variable that was not controlled for in corpus construction. Consequently conclusions drawn from the dataset need to account for this uncontrolled factor. Besides uncontrolled variables another limitation of the presented dataset is the simplified interpretation of Wikipedia talk page discussions. The text of each discussion is reconstructed to its state at one single point in time. Given this text the discussion is viewed as a chronological sequence of utterances that reply to one another, similar to a face-to-face conversation between people in real life. The problem is that single utterances can be changed or deleted even after other editors responded to them. This loose structure makes Wikipedia deliberations hard to comprehend in entirety, because every single change to the discussion text needs to be reviewed in order to understand the true progression of the complete deliberation. As a result of this structure the two dimensions of time and text progression are not aligned in Wikipedia talk page discussions (See figure 4.2). There is no equivalent to so structured discussions in natural conversation and human interaction. Presumably this is why most chats and online forums structure written conversations as chronologically ordered and immutable sequence of utterances or posts to mimic the nature of actual conversations, where time and discussion progression parallel each other (See figure 4.3). This difference in conversation structure makes it difficult to apply conclusions deferred from the presented dataset to deliberation in general. To estimate if this simplification fails to capture the actual deliberation and how much this hampers transferable conclusions it needs to be analysed if the Wikipedia guideline suggestion to use talk pages like a chronological ordered chats is followed in most deliberations or not. Similar simplifications

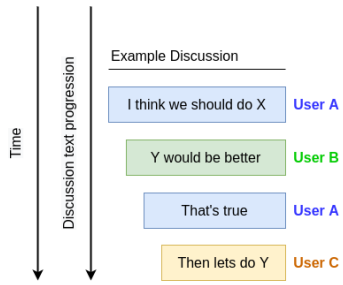


Figure 4.3: Schematic depiction of the progression of a chronological ordered deliberations, where new utterances can only be added below the last and not get changed.

of Wikipedia talk page discussions premise most research on this topic and the pitfalls that come with it get hardly mentioned. Therefore we think that the suggested analysis is important to asses the value of future research on Wikipedia talk page discussions with the goal to find patterns that apply to deliberation in general.

Chapter 5

Conclusion

This last chapter shortly summarizes the presented work and highlights our key findings. Given the goal to understand the relation between argumentative strategies and the success of deliberations we start by stating what is meant by argumentation strategy and when we view deliberation as successful:

(1) As presented in [Al-Khatib et al., 2018a] for a deliberation that consists of subsequent utterances every utterance can be labeled with a category that describes its argumentative function inside the deliberation. The resulting sequence of these labels is the argumentation strategy of the deliberation.

(2) A deliberation is interpreted as successful if the participants reached agreement towards the end. If the deliberation ends without consensus it is viewed as failed.

In order to examine the impact that these strategies have on deliberation success we present an approach for creating a dataset of deliberations labeled as successful or failed from discussions on Wikipedia talk pages while controlling the discussion subject. To find successful and failed deliberations the *Requests for comment* process on Wikipedia is leveraged where discussions that failed to reach consensus get resolved in a second deliberation with impartial editors. As Wikipedia does not provide a link between such a failed predecessor discussion and the RfC discussion that resulted from it the main piece of work is to join these pairs. For this purpose a pipeline is presented that finds 421 RfC-predecessor pairs from 17 thousand RfC deliberations available on Wikipedia. Attempts to train classifiers with this dataset failed and the corpus could not be extended further. The obtained dataset of both 421 successful and failed deliberations is expected to be free of topical bias between the success and the failure subset due to our creation process. Therefore it can be used to draw meaningful conclusions about deliberation success. A first analysis of differences between the two subsets shows that in successful deliberation editors refer to the opening statement more often. We conclude that these

results highlight the importance of meaningful and clear opening statements for goal oriented deliberation.

Bibliography

Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. Modeling deliberative argumentation strategies on Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555, Melbourne, Australia, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1237. URL https://aclanthology.org/P18-1237.

Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. Modeling Deliberative Argumentation Strategies on Wikipedia. In Iryna Gurevych and Yusuke Miyao, editors, *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 2545–2555. Association for Computational Linguistics, July 2018b. URL https://www.aclweb.org/anthology/P18-1237/.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL https://arxiv.org/abs/2004.05150.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1305. URL https://aclanthology.org/D18-1305.

Jane Im, Amy Zhang, Christopher Schilling, and David Karger. Deliberation and resolution on wikipedia: A case study of requests for comments. volume 2, pages 1–24, 11 2018. doi: 10.1145/3274343.

Keith Maki, Michael Yoder, Yohan Jo, and Carolyn Rosé. Roles and success in Wikipedia talk pages: Identifying latent patterns of behavior. In

Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1026–1035, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1103.

Lu Wang and Claire Cardie. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2113. URL https://aclanthology.org/P14-2113.