

# **Analyzing Controversy in Wikipedia**

by

Hoda Sepehri Rad

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

© Hoda Sepehri Rad, 2015

# Abstract

This thesis describes a novel controversy model that helps the current manual process in automatically identifying controversial Wikipedia articles and warning readers about disputable information contained in these articles. The model is based on identifying collaboration patterns among editors and inferring their attitudes towards one another. These are exploited in the form of a social network representing the overall structure of history of collaborations of editors of each article. A set of features, rooted at sound theories of social behavior, are extracted from each network to distinguish controversial articles from other articles.

To infer attitudes, a novel supervised approach is employed based on votes cast in Wikipedia admin elections. The combination of structural features extracted from each network, and the method for inferring attitudes of editors provides an accurate and efficient controversy model as demonstrated by several experiments.

Also, a systematic evaluation and comparison of previous controversy models is provided. The results show the inefficiency of most of these models in capturing the complex process of formation of controversy, and express the power of editors collaboration networks for modeling this process.

Finally, to give more insight about controversial topics, two different approaches are proposed to analyze controversy at a fine-grained level. The first approach aims to separate the most controversial parts of each article from other non-controversial and reliable parts. This approach is shown to be a challenging problem due to both designing a suitable method and providing a quantitative evaluation. On other hand, the second approach helps to get a ranked list of the revisions that contributed most to controversy of the article. For this approach, a solution based on maximum coverage problem is proposed and its usefulness is shown by quantitative results.

# Preface

Most of the research presented in this thesis has benefited from guidance of my supervisor, Professor Denilson Barbosa. He also has assisted greatly in writing the manuscript by providing editorial feedback.

Chapter 1 and Chapter 2 are my original work.

The main part of Chapter 5 and some parts of Chapter 3 have been published in 23rd ACM conference on Hypertext and social media [68]. In this work, I and Aibek Makazhanov are the main contributors. Aibek contributed by designing and implementing the preliminary ideas of building PV networks, which later on were corrected, improved, and expanded by me. I also designed experiments of this part, and was the key role in analyzing the results. Aibek was responsible for implementing them and assisted in analyzing the results. The general concept of collaboration networks and the ideas discussed in Chapter 3 are designed and developed solely by me. I wrote the whole manuscript as well. Denilson Barbosa and Davood Rafiei both assisted by discussing ideas and editing the manuscript.

A more complete version of the above paper was prepared and extended by me, which has been published in ACM Transaction Intelligent Systems and Technology. In this work, I provided more analytical examples, proofs and experiments. I also wrote the manuscript. Denilson Barbosa assisted by editing the manuscript and providing responses to reviewers comments.

Finally, I am the sole contributor of ideas proposed in Chapters 4, 6 and 7. Chapter 4 has been published in 8th International Symposium on Wikis and Open Collaboration, while Chapter 6 has appeared in part in the 20th International Conference on World Wide Web. Denilson Barbosa assisted in writing the manuscripts of both of these publications.

To my mom who did not let me to “give in without a fight”.

*Allah will judge between them on the judgment day about that wherein they are differing.*

– Holy Quran- 2: 113.

# Acknowledgements

First and foremost, I would like to thank “Allah”, the most Beneficent, the most Merciful, whose many blessings have made me who I am today. I’m also very thankful to “Allah” for giving me very supportive and loving family. My husband, Majid, had a big role in my success in this long path, and without his love, helps, and sacrifice this work would not have been possible. He bore living with a mostly busy, and sometimes frustrated graduate student for 7 years, and gave lots of assistance and support in taking care of our son in the last 1.5 years of my study. My parents and my two wonderful sisters also had a big in impact in this work by giving unconditional love, constant support and encouragement. My son, “Yaseen” has been a blessing in my life and his birth permanently changed my view towards life. He has been the best stress reliever in all those tough times I had in progressing this work, and his bright smile always reminded that nothing is such a big deal in life! I’m very thankful to all of you my dearests.

In the academics, the first and most important person I would like to thank is my supervisor, Denilson Barbosa. He was always very supportive, patient, and encouraging. I could have not asked for a better supervisor in this regard. He gave me lots of freedom and opportunities to explore different ideas and solutions and provided me with great insights and advice when necessary. I learned a lot from him during these years. In particular, his passion for perfection in English and clear writing, significantly improved my communication skills. His confident character and encouraging optimism were also very inspiring for me. A big thanks for all of these Denilson.

I would also like to extend my appreciation to my supervisory committee members, Osmar Zaiane and Dale Schuurmans for their great feedback and comments

on my seminars and thesis drafts. I'm also grateful to Aibek Makazhanov and Davoud Rafiei, who both had an important contribution in this thesis. Aibek is the most hard-working student I have ever seen, and he genuinely spent a lot of time and effort to implement some of the ideas discussed in this thesis. Davood also gave many constructive feedback and ideas for improving our joint work. It was a great pleasure to collaborate with both of you. I also benefited from discussions about modeling controversy in Wikipedia and some implementation issues with Taha Yasseri, and Shiri Dori, both great researchers in their fields. I would also like to thank Johannes Daxenburger for providing us with labeled data from his edit-classifier.

Finally, I have had a chance to see some of my best friends in life since the beginning of my stay in Edmonton. I'm specially thankful and proud for being a member of our "gang" group friends, with whom I have a lot of good memories. I'm also grateful to Sabereh Rezayi, Leila Zargar Zadeh, Leila Shafiei, Neshat Pazooki and Elham Taghavi Mehr for being supportive, helpful and honest friends. A special thank also goes to our "Secord family", Maedeh RoodPeyma and Behzad Vafaeian for being there "whenever" we needed any help.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background on Wikipedia . . . . .	1
1.1.1	Editing Process . . . . .	1
1.1.2	Types of Articles . . . . .	3
1.2	Motivation and Goals . . . . .	5
1.3	Thesis Statements . . . . .	7
1.4	Overview of Proposed Methods . . . . .	7
1.5	Summary of Contributions and Published Results . . . . .	9
1.6	Thesis Organization . . . . .	10
<b>2</b>	<b>Related Work</b>	<b>11</b>
2.1	Trust and Quality Assessment in Wikipedia . . . . .	11
2.2	Visual Analysis of Collaborations in Wikipedia . . . . .	14
2.3	Controversy Modeling in Wikipedia . . . . .	15
2.3.1	Article-level Analysis . . . . .	16
2.3.2	Fine-grained Analysis . . . . .	18
2.3.3	Evaluation of Proposed Methods . . . . .	19
2.4	Opposing Views in Other Social Media . . . . .	20
2.4.1	Supervised Methods . . . . .	20
2.4.2	Unsupervised Methods . . . . .	21
<b>3</b>	<b>Identifying Controversial Articles Using Collaboration Networks</b>	<b>24</b>
3.1	Proposed Method . . . . .	25
3.1.1	Overview . . . . .	25
3.1.2	Collaboration Network . . . . .	25
3.1.3	Structure Classifier . . . . .	26
3.2	Experimental Results . . . . .	29
3.2.1	Dataset . . . . .	29
3.2.2	Comparison with other methods . . . . .	30
3.2.3	Classification Accuracy . . . . .	31
3.2.4	Feature Analysis . . . . .	33
3.2.5	Effect of History Length . . . . .	35
3.2.6	Filtering Networks by Active Editors . . . . .	37
3.3	Conclusion . . . . .	38
<b>4</b>	<b>Comparative Study of Other Controversy Models</b>	<b>39</b>
4.1	Examined Methods . . . . .	40
4.1.1	Mutual Reverts . . . . .	40
4.1.2	Bipolarity . . . . .	42
4.1.3	Basic REA . . . . .	43
4.1.4	Structure classifier . . . . .	44
4.1.5	Meta classifier . . . . .	45



4.2	Evaluation Framework . . . . .	46
4.2.1	Classification vs. Ranking . . . . .	47
4.2.2	Metrics . . . . .	48
4.2.3	Dataset . . . . .	48
4.3	Experimental Results . . . . .	49
4.3.1	Discrimination Power . . . . .	49
4.3.2	Cost of Training . . . . .	52
4.4	Categorization of Controversy Models . . . . .	54
4.5	Conclusion . . . . .	56
<b>5</b>	<b>Building PV Collaboration Networks</b>	<b>57</b>
5.1	Extracting Contributing Editors . . . . .	60
5.2	Identifying Collaboration Relations . . . . .	60
5.3	Classifying Collaboration Relations . . . . .	61
5.4	Building Collaboration Profiles . . . . .	62
5.4.1	Individual Features . . . . .	63
5.4.2	Directional and Mutual Features . . . . .	66
5.5	Classifying Collaboration Profiles . . . . .	67
5.5.1	Leveraging Admin Elections . . . . .	67
5.6	Experimental Results . . . . .	68
5.6.1	Dataset Description . . . . .	68
5.6.2	Overall Prediction performance . . . . .	69
5.6.3	Comparison with other methods . . . . .	71
5.7	Conclusion . . . . .	72
<b>6</b>	<b>Fine-grained Analysis of Controversy at Text-unit level</b>	<b>73</b>
6.1	Problem Formulation . . . . .	74
6.1.1	Computational Complexity . . . . .	75
6.2	Defining Contribution Function . . . . .	76
6.2.1	Computing with Revision History . . . . .	77
6.2.2	Assessing Current Controversy Models . . . . .	78
6.3	Evaluation . . . . .	80
6.3.1	Difficulties of a Human Judgment Experiment . . . . .	81
6.4	Conclusion . . . . .	82
<b>7</b>	<b>Fine-grained Analysis of Controversy at Revision level</b>	<b>84</b>
7.1	Problem Formulation . . . . .	85
7.2	Coverage-based Contribution Function . . . . .	86
7.2.1	Background . . . . .	86
7.2.2	Adapting to Our Problem . . . . .	87
7.2.3	Defining Term-global Score . . . . .	88
7.2.4	Defining Term-Local Score . . . . .	89
7.2.5	Summary of the proposed Revision Selection Method . . . . .	90
7.3	Evaluation . . . . .	92
7.3.1	Set-level Evaluation . . . . .	92
7.3.2	Individual-level Evaluation . . . . .	93
7.4	Comparison with Other Methods . . . . .	95
7.5	Experimental Results . . . . .	96
7.5.1	Set-level Results . . . . .	97
7.5.2	Individual-level Results . . . . .	98
7.5.3	Detailed Examples . . . . .	99
7.6	Conclusion . . . . .	103

<b>8</b>	<b>Conclusion and Future Work</b>	<b>106</b>
8.1	Conclusion . . . . .	106
8.2	Future Works . . . . .	108
8.2.1	Improving Attitude Inference Model . . . . .	108
8.2.2	Modeling Controversy in other Domains . . . . .	109
8.2.3	Representation of Selected Revisions . . . . .	109
8.2.4	Considering Other Factors in Selecting Revisions . . . . .	110
8.2.5	Other approaches for fine-grained Analysis of Controversy .	110
	<b>Bibliography</b>	<b>111</b>
<b>A</b>	<b>Appendix</b>	<b>119</b>
A.1	Unit-level Analysis assuming Independent Units . . . . .	119
A.2	Mutual Reverts is a Monotone Score . . . . .	119

# List of Tables

3.1	Results of identifying controversial articles . . . . .	32
3.2	Accuracy results of each method on different ranges of the edit history	36
4.1	Summary of the main characteristics of the studied methods . . . . .	40
4.2	Statistics of datasets used for comparative study of controversy models . . . . .	49
4.3	Comparison of the studied methods in terms of accuracy, precision, and recall . . . . .	51
5.1	Statistics of election dataset . . . . .	69
5.2	Statistics of the extracted, and the mapped election data . . . . .	69
5.3	Results of predicting votes from collaboration profiles . . . . .	70
5.4	Top-15 important features of vote classifier in predicting votes . . . . .	71
7.1	List of top-20 terms in “Abortion” article . . . . .	88
7.2	Results of comparison of methods in terms of MAP . . . . .	99
7.3	Top-10 selected revisions for Abortion article . . . . .	103
7.4	Top-10 selected revisions for Osama bin Laden article . . . . .	104

# List of Figures

1.1	Difference between two revisions using a diff tool . . . . .	2
1.2	Example of a dispute tag . . . . .	5
2.1	Example of applying WikiTrust [1] . . . . .	13
2.2	Example of visualization of collaborations of editors [40] . . . . .	15
3.1	Overview of our proposed method for identifying controversial articles . . . . .	25
3.2	Eight different triad types used in our Structure classifier . . . . .	29
3.3	Analysis of the top-10 structural features . . . . .	34
3.4	Effect of limiting history length . . . . .	36
3.5	Effect of filtering collaboration networks by top active editors . . . . .	38
4.1	Distribution of articles in our dataset in terms of baseline methods . . . . .	50
4.2	Effect of training size on accuracy . . . . .	53
5.1	Partial edit history of article on Anarchism. . . . .	58
5.2	The workflow of building collaboration networks in our work . . . . .	59
5.3	Distance distribution of revision pairs editing the same section . . . . .	61
5.4	Three groups of features used in building the collaboration profile . . . . .	63
5.5	Tag clouds of top agreement and disagreement selected comment terms . . . . .	64
6.1	Example of obtaining modified revisions under the inclusion model . . . . .	78
6.2	Monotonicity test for the studied methods . . . . .	79
7.1	An example of calculating term-global and term-local scores. Inserted, deleted and unchanged terms are shown in green, red and blue respectively. . . . .	91
7.2	MR-drop of different selection methods at different numbers of selected revisions . . . . .	98
7.3	Precision at $k$ of different selection methods measured by using edit-classifier of Daxenberger et al. [15] . . . . .	99

# Chapter 1

## Introduction

### 1.1 Background on Wikipedia

Wikipedia<sup>1</sup> is a collaboratively edited online encyclopedia, written by volunteers around the world. It is the largest and one of the most successful encyclopedias in the world and currently contains 30 million articles in 287 languages, including over 4.5 million in English [29]. According to statistics reported by New York Times, in February 2014, Wikipedia ranked fifth globally among all websites, having over 18 billion page views per month [33]. While anecdotal evidence points to problems and virtues in relying on Wikipedia [14], the trend seems to be that Wikipedia will indeed become the primary source of reference for most common knowledge in the world. In the next sections, we give more background about this popular knowledge-base.

#### 1.1.1 Editing Process

Wikipedia is based on an open-access model, where anyone can edit its articles or create new ones. For most articles, editors even do not need to have an account to edit text of articles. These *anonymous editors* that are identified by their IP address might have less authority compared to *registered editors*. For instance, in the English edition of Wikipedia, anonymous editors cannot create new articles.

Editors should work together to write articles collaboratively, where content and style of articles are decided based on consensus. Wikipedia has some tools for

---

<sup>1</sup>Wikipedia.org

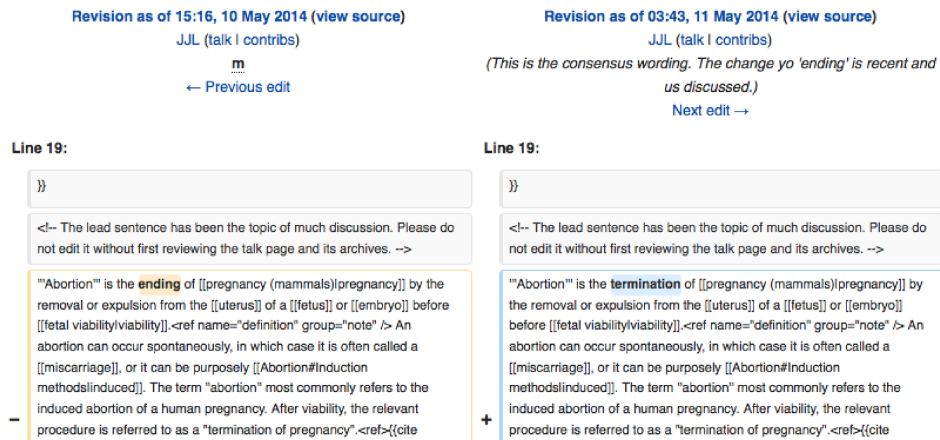


Figure 1.1: Difference between two revisions using a diff tool

facilitating this collaboration. For instance, Wikipedia records all changes applied to articles in the form of a *revision history*, where each *revision* corresponds to a particular state of an article. This state contains a set of modifications applied to a version of the article by an editor at a specific time and date. Editors can use the revision history to view older versions of the article or more importantly compare different revisions line by line using a diff tool [35]. An example of this comparison is shown in Figure 1.1.

In addition, as all changes are saved in the revision history, editors can undo undesirable changes and restore lost content by *reverting* the article to an earlier version. In particular, reverts are frequently used to fight *vandalism*, which according to Wikipedia is any deliberate attempt to damage the integrity of articles. Examples of vandalism are mass deletes, adding irrelevant or vulgar material, and inserting obvious nonsense to the article.

Another tool for aiming collaboration of editors is the *discussion page* (a.k.a the *talk page*) associated with each article. In these pages, editors can exchange ideas and talk about different issues related to the article to reach consensus and coordinate the work among themselves.

Wikipedia also enforces special policies to protect articles against abuse and conflicting cases. One of the earliest and most important example of these policies is the *Neutral Point of View (NPOV)* which states that all articles should be written

from a neutral point of view, representing significant views fairly, proportionately and without bias. In this way, when disagreement between editors is inevitable on certain polarized topics, this policy helps editors to accept all encompassing arguments instead of arguing on what is right and what is not.

In addition to these tools and policies, there is a group of privileged editors named *admins* who have special power to manage the collaboration of editors. For instance, admins have the ability to delete articles (i.e. in case the article does not comply with Wikipedia standards and policies such as advertising a company), lock articles from being changed in case of vandalism, and block editors from editing. Admins are selected through *elections* where a trusted, good-standing editor requests for adminship and other (registered) editors cast positive, negative or neutral votes towards the promotion of the candidate editor. Candidates that are successful in their elections will be granted admin status, where they mostly work on coordination and conflict resolution tasks rather than direct contributions to articles.

Giving positive votes in admin elections is not the only way of showing support and appreciating the work of other editors. Another way for this purpose in Wikipedia is through *barnstars*, which are star-type icons that editors can use to reward other editors for their hard work and efforts. These awards are put on the page of the awarded editor with an explanation of why they are given. This helps other editors and reader to see and get informed about these awards.

### **1.1.2 Types of Articles**

The open access model of Wikipedia, which contrasts to expert-driven approach of traditional encyclopedias, has attracted many contributed world wide and has been one of the main reasons of its success. In fact, according to a study in the journal Nature, scientific articles in Wikipedia have comparable quality to similar entries in the traditional Britannica encyclopedia [29].

However, this open access model also comes with the risk of vandalism and publication of inaccurate information that lead to have unreliable articles or articles where some of their revisions are untrustworthy. To overcome this weakness, Wikipedia uses an internal quality assessment system, which relies on judgment

of the editors community. The range of this assessment system starts with “stub” articles as the lowest quality articles and continues with classes “C”, “B”, “A” and “good articles”. After that, there is the class of *featured articles*, which are the highest quality articles. Before an article becomes “featured”, it goes through a detailed review process done by reviewers, who are selected from experienced editors. In this process, factors such as accuracy, neutrality, completeness, and writing style are assessed. Currently, of more than 4 million articles in English Wikipedia, about 0.1% of them are classified as featured articles [29].

The concern with quality and trustworthiness of articles is especially important for articles whose topics fall in a naturally polarizing category such as religion, history, or politics. For these topics, different viewpoints and opinions exist, which sometimes make it difficult to know where the truth lies. These different and possibly opposing viewpoints can cause editors of corresponding articles to take different sides and argue with each other on what should be included in the article. When these arguments get serious and are not managed correctly, they can cause articles to become biased towards a specific viewpoint and not adhere to the NPOV policy of Wikipedia. They also can cause articles to experience one or more *edit wars* among editors in their revision histories. In edit wars, two or more editors repeatedly undo each other’s work, rather than trying to resolve the conflict rationally.

To help manage these conflicts and to warn readers and editors about the disputed state of these articles, Wikipedia has specific templates that can be put in the beginning of articles or part of them (like a specific section). These templates are referred to as *dispute tags* and are designed to cover different issues and scopes related to controversy. For instance, `{{totally-disputed}}` and `{{POV}}` are related to disputed content and violating NPOV policy at the article level respectively, while `{{disputed-section}}` is a tag for marking a section as being disputed. Figure 1.2 shows an example of a dispute tag when readers see an article with such a tag. A list of some of the past known *controversial articles* is also kept by Wikipedia community<sup>2</sup>, so that editors can check them time to time as these articles are likely to suffer from future disputes.

---

<sup>2</sup>[http://en.wikipedia.org/wiki/List\\_of\\_controversial\\_articles](http://en.wikipedia.org/wiki/List_of_controversial_articles)



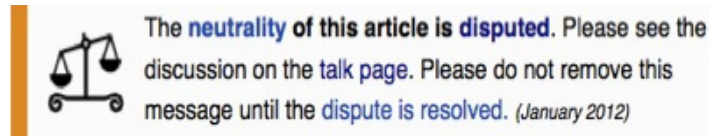


Figure 1.2: Example of a dispute tag

## 1.2 Motivation and Goals

The current set of controversial articles (manually tagged by editors) forms a small fraction of all articles. However, it should be noted that these articles span a wide range of topics, including many well-known and popular topics (thus attractive to all Wikipedia readers). For instance, about 9% of the top 1000 visited articles according to statistics obtained in 2015 are among the list of controversial articles [31]. In addition, there might be several other controversial and disputed articles that have not been tagged by its editors.

In this thesis, we are interested in automatically identifying articles that contain a history of dispute and controversy. This can improve the current manual process, which relies on editors to tag articles manually. It can also be useful as a pre-processing tool for analyzing topics and patterns of collaboration among editors that lead to controversy. The need for such tools can be seen with the growth of problems related to bias and pushing points of views in Wikipedia studied in some recent works [7, 22]. Automatic and objective tools can also replace the possible biased and guided editing process, where some editors might even resist against putting dispute tags.

Moreover, the current tools and methods in literature do not provide any information about controversial articles in Wikipedia beyond specifying them as being controversial. What is more important than knowing whether an article is controversial or not, is the ability to know what has caused the article to be controversial and what are the main discussed issues. Hence, we also consider a fine-grained analysis of controversy to help readers better understand issues and opposing views expressed in controversial articles.

For this analysis, we try to highlight the text-units such as paragraphs or sections of each controversial article where most dispute and conflict between editors hap-

pened in them. This is because controversy in an article may involve disputes about several independent issues, each of which covering a different aspect of the main topic. Identifying these specific controversial parts is not only useful to understand what the dispute is about, but also helps to identify which parts of an article should be considered with caution, and which parts can be still relied on. For instance, the article “Holy sites in Islam” was one of the most controversial Wikipedia articles [76], which later on was replaced by two separate articles based on viewpoints of the two sects of Islams (i.e. Shia and Sunni). In the original article, it can be seen from the discussion page and the history of some of revisions that the controversy was restricted to the part mentioning “Al-Aqsa Mosque” as the third holiest site of Muslims. Knowing that naming this site was the main reason of controversy can help readers to judge this part of the article and its other parts better.

As another way of providing a fine-grained controversy analysis, we aim to find the revisions in the history of each article that are responsible for making the article controversial. Extracting these revisions out of a large pool of revisions that are mixed of peaceful (i.e. such as fixing an error, or adding new information), vandalism and disputed revisions provides a fast way for editors and readers to get informed about the most important conflicting points. For instance, in “Abortion” article, there is a history of conflict on the issue of a “the existence of a link between abortion and breast cancer”. Finding revisions that edit the article in regard to this issue can help to see one of the reasons that led this article to become controversial, along seeing possible related biases and viewpoints.

Our gold standard for controversial articles consists of the list of past known controversial articles that as explained before is kept by the Wikipedia community. These articles, according to Wikipedia’s policies, “ regularly become biased and need to be fixed, or are articles that were once the subject of a neutral point of view dispute and are likely to suffer from future disputes.”<sup>3</sup>. It should be noted that while most of the articles in this list are about controversial topics, an article covering a controversial topic is not necessarily a controversial article. This is because it might be the case that editors have managed to collaborate effectively and write the article

---

<sup>3</sup>[http://en.wikipedia.org/wiki/List\\_of\\_controversial\\_articles](http://en.wikipedia.org/wiki/List_of_controversial_articles)

in an unbiased form that covers all opposing viewpoints, and therefore the article will not become controversial.

### 1.3 Thesis Statements

In this thesis, we are interested in answering the following research questions:

- Can we accurately predict whether or not a Wikipedia article is controversial by analyzing the way in which editors interact?
- Can we determine which specific text-units (e.g., sections or paragraphs) within a controversial article contribute the most to the controversy of the said article?
- Can we determine which revisions (i.e., changes applied by an editor at a specific time) of a controversial article contribute the most to the controversy of the said article?

The first question helps to identify controversial articles and automate the current manual process as discussed in previous section. For this purpose, we analyze articles at the *whole-article level*. On the other hand, the later two questions help to give more insights about controversial articles by locating the sources of controversy at the *text-unit level* or *revision level*. For these two questions, we analyze articles at a fine-grained level that was not addressed in previous work.

### 1.4 Overview of Proposed Methods

Wikipedia is a collaborative system with many different types and levels of social interactions among its contributors. One of the main type of interactions between editors is collaboration of editors on writing articles. This type of interaction is crucial information for characterizing Wikipedia articles.

Using these collaboration interactions, we give a very effective method for detecting controversy in Wikipedia. Our method is based on understanding *collaboration patterns* among editors and inferring their *attitudes* towards one another.

In particular, we employ these inferred attitudes in the form of a network structure representing collaboration patterns of each Wikipedia article. The network of collaborations for each article consists of its main editors connected with edges having positive or negative signs, denoting attitudes of editors. We analyze these *collaboration networks* using a set of *structure features* that are commonly used in analysis of social networks. We then use these features to train *Structure classifier*, which labels articles as either controversial, or non-controversial.

Structure classifier is a general model that can be applied on different types of collaboration networks. We propose a novel method for building these networks by utilizing information in Wikipedia admin elections. This type of collaboration networks, referred to as PV (Profiles and Votes) networks, results in distinguishing the two classes of controversial and non-controversial articles with very high accuracy. It also provides us with a powerful controversy model that works well even with having access to only a limited part of history of articles or with articles having short history.

To build *PV networks*, we exploit the admin election repository as a source of social interactions to infer attitudes of editors. We found that there is a strong correlation between how editors vote for each other, and previous history of their collaborations. We utilized this correlation by training a classifier that uses signs of votes as its training labels and a concise form of history of collaborations of voters and candidates as its training instances. This concise form, referred to as *collaboration profile*, is built by using an extensive set of features from individual and pairwise edit activities of collaborating editors.

Finally, we analyze controversy at two fine-grained levels of *text-units*, and *revisions*. We formulated both of these analyses using an optimization framework, where we show that it requires its objective to satisfy the two properties of *submodularity* and *monotonicity* due to its computational complexity.

For text-unit level analysis, we discuss how current article-level controversy models can be used as a candidate objective function, and examine these models with respect to these desired computational properties. The results show that these models are not suitable candidates for this problem. We also, consider different

ways for evaluation and discuss shortcomings of them.

For revision-level analysis, we propose an objective function based on *maximum coverage problem*, which is a well-known problem in approximation algorithm theory. We show that our method selects better revisions compared to baselines and methods adopted from literature, while satisfying the desired computational properties.

## 1.5 Summary of Contributions and Published Results

Our main contributions are as follows:

- We build an effective model for identifying controversial articles in Wikipedia based on inferring attitudes of editors and building a network of their collaborations that led to a significant improvement over previous methods. We proposed this method first in the 23rd ACM conference on Hypertext and social media (HyperText'12) [68], and later presented it in more details in ACM Transactions on Intelligent Systems and Technology (TIST) [67] by including more experiments and analyses.
- We show a novel approach for learning to infer attitudes of editors using admin elections and an extensive set of features, summarizing activities and interactions of editors. This contribution is based on a joint work with other collaborators that was published in HyperText'12 [68].
- We examine current proposed controversy models under a standard evaluation framework, for the first time, to fill the gap of lack of comparison and systematic evaluation that existed in this area. The results of this part were published in the 8th International Symposium on Wikis and Open Collaboration [66].
- Finally, we propose and study two novel problems related to analyzing controversy beyond the whole article-level to help giving more insight on what the controversy is about. The preliminary results of text-unit analysis including the introduction of the problem and possible challenges for its evaluation

were presented in the 20th International Conference on World Wide Web (WWW'11) [65].

## **1.6 Thesis Organization**

The remaining of this thesis is organized as follows: in Chapter 2, we briefly review some of the work related to studying controversy and similar issues in Wikipedia, and in other social domains. In Chapter 3, we describe our method for identifying controversial articles using collaboration networks and Structure classifier. In Chapter 4, we study this model in more details and compare it against four other controversy models from the literature under different settings. In Chapter 5, we present our method for building PV networks, and also explain the extensive set of features we extract to build collaboration profiles of editors. In Chapters 6 and 7, we propose our work on analyzing controversy at two levels of text-units and revisions respectively. Finally, in Chapter 8, we conclude this thesis and point out to some future directions.

# Chapter 2

## Related Work

Our work is related to works in several different areas as we explain briefly below.

### 2.1 Trust and Quality Assessment in Wikipedia

There is a large body of work related to trust and quality assessment in Wikipedia. In these works, typically a trust score is assigned to an article [8, 34, 83], to selected parts of an article [1], to each revision of an article [19], or to its editors [2, 12, 36]. These works often use information from the edit history of articles, including edit actions and the way they evolve in response to edits. For instance, *reverts* (i.e. undoing an edit) and *restores* (i.e. changing back to an earlier revision) are treated as direct indications of distrust and trust in most works.

Priedhorsky et al. [62] had access to visiting statistics of articles and used the number of times a revision of an article has been visited as a notion of quality and impact of a contribution. They showed that what readers mostly see when they visit Wikipedia articles come from edits of the top frequent editors who are responsible for most of the contributions. The authors also studied the impact of damages and edits that compromise quality of articles, and found out that the probability of a typical view encountering a damage is very low, however this probability has been increasing over time. Furthermore, the authors classified different types of damages and analyzed their impact on readers. For instance, they showed that mass deletes or insertion of nonsense into an article are quite easy to detect, and have not much impact on readers compared to adding false information or advertisement and spam

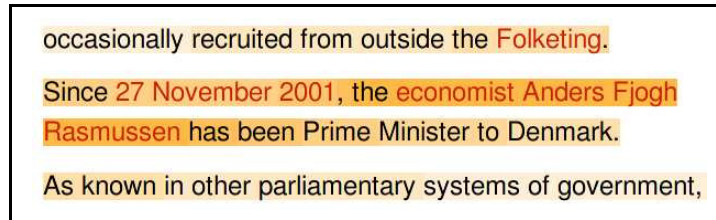
information.

Other features that were used to establish some notion of trust are reputation of editors of previous revisions [1, 19, 83]. Zeng et al. [83] developed a model for assessing trustworthiness of each revision of a Wikipedia article. Their model is based on a Bayesian network, where trustworthiness of a particular revision is a function of the trustworthiness of its previous revision, reputation of the editor of the previous revision, and the amount of text added or removed between the previous revision and this current revision. Reputation of editors is calculated by categorizing editors into four groups of admins, registered users, anonymous editors, and blocked editors and assuming a random beta distribution for trust values of each group. In this way, each editor is assigned a constant reputation value according to the group he belongs to.

In contrast to Zeng et al., Adler et al. [2] developed a reputation system assigning dynamic reputation values to editors that depend on their actions. In their method, editors gain reputation when their edits are preserved in subsequent revisions, and they lose reputation when their edits are reverted or undone. The authors later extended their work by using this developed reputation system in assigning trust values to each word in each revision of a Wikipedia article [1]. Trust values are computed based on reputation of the original editor of each word, as well as the reputation of all editors who edited the text in proximity of that word. The authors also developed a novel visualization technique showing trust values of different words of a specific revision of an article using varying text-background colors to help readers distinguish the trustworthy parts from low quality and unreliable parts. An example of this type of visualization is shown in Figure 2.1. Similar trust assessment and editors' reputation systems based on the idea of text stability can be seen in later works by other authors [12, 36, 42].

Another approach for assessing quality and trustworthiness of articles is to use a combination of statistics such as the number of authors, the number of in-links and out-links, length of the article, and other similar statistics. in a supervised machine learning framework. The high number of unique editors, long text, and high number of revisions are among the most important factors distinguishing high





occasionally recruited from outside the Folketing.  
Since 27 November 2001, the economist Anders Fjogh Rasmussen has been Prime Minister to Denmark.  
As known in other parliamentary systems of government,

Figure 2.1: An example of colouring words by their trust values after an edit that changes the name of prime minister from “Fogh” to “Fjogh” [1]

quality articles from low quality articles found in different studies [26,49,81].

The gold standard in trust and quality assessment work is to have featured articles and good articles as high quality and trustworthy articles, and consider less quality and trust values for classes B, C, and stub articles. Comparing high quality articles with low quality articles can be done at the whole revision history of articles when usually the task is a binary or multi-class classification. It can also be done in a dynamic way and at revision-level, where the evolution of the trust score is considered [36]. Also, editors’ trustworthiness and reputation modeling works usually resort to verifying calculated scores for admins and vandals as these editors are assumed to have high and low reputation respectively. In addition, showing a strong correlation between previous calculated reputation of editors and their future behaviour is another evaluation paradigm used in reputation modeling works [2,36]. For instance, those editors who were found to have a high reputation based on the proposed models tend to have more long lasting contributions in their future edits, while those with low reputation were found to contribute to edits that were mostly reverted.

In this thesis, we study “controversy” , which is a different concept than trust and quality. Controversy arises when sufficiently different, and often contradictory views about a subject exist. In these cases, it hard or impossible to judge where the truth lies. While it is reasonable to label controversial articles as untrustworthy, the converse is not necessarily the case: there are many reasons that make an article untrustworthy, such as vandalism or the presence of incorrect information. Besides, trust and quality values can be measured at the revision or at the article levels. Controversy, on the other hand, is the result of serious disagreement in opinions

between two or more people over a *prolonged period of time*<sup>1</sup>, and therefore should be measured over a sequence of actions and edits and is not applicable for a single revision.

## 2.2 Visual Analysis of Collaborations in Wikipedia

Our work is also related to research on visualization of edit history and the collaboration process of Wikipedia articles. The aim of these techniques is to help readers to grasp a high-level information about the general characteristics of the articles. For instance, an observation from these works is that different structure can be found for controversial articles compared to other articles. In this regard, Suh et al. [72] developed a visualization technique where editors are connected with disagreement relations that are extracted using revert actions. In this visualization, editors are rearranged using a special layout so that the distance between them will be proportional to the degree of their disagreement.

Brandes et al. [9] considered short time difference between consecutive revisions as an indicator of disagreement and proposed a visualization technique showing the dominant editors and roles of editors such as “reviser” or “being revised”. They also proposed another technique for visualizing network of editors of each article. Based on their analysis, they found out that controversial articles tend to be more similar to bipolar graphs compared to other types of articles [8].

Also, some authors tried to partition and find communities of editors agreeing with each other and disagreeing with other communities [5, 40]. Figure 2.2 shows an example of this partitioning done by Kittur et al. [40].

However, these visualization techniques are only useful for grasping editors interaction qualitatively. For instance, knowing the perspective and opinion of each of the four cluster of editors in Figure 2.2 is not possible without manual extraction of edits of editors and background knowledge about the topic of the article. Moreover, Wikipedia articles usually have a long history of edits, containing thousands of revisions, and constantly gets updated, which make application of visualization

---

<sup>1</sup>C.f. the Oxford English Dictionary.

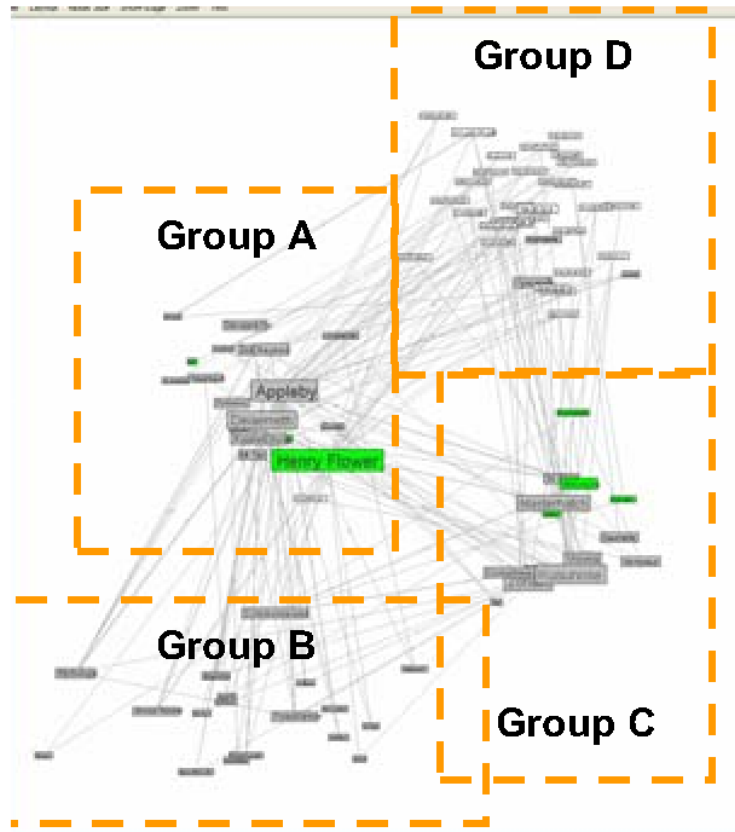


Figure 2.2: An example of visualization of collaborations of editors in a Wikipedia article [40]

techniques for this domain to become limited.

## 2.3 Controversy Modeling in Wikipedia

The importance of analyzing controversy in Wikipedia has been pointed out by other researchers in recent years. For instance, Flöck et al. [22] studied the necessity of developing tools and procedures to remove barriers for adding legitimate, balanced and unbiased representation topics in Wikipedia articles. They also discussed several evidences of systematic bias and resistance on accepting new view points. Examples of these evidences are the start of projects such as Conservapedia (a wiki written by former Wikipedia editors claiming that Wikipedia is biased towards liberalism), and the highly skewed distribution of contributors in terms of geography, age, and gender. For instance, according to Wikipedias project page on

Systemic Bias [30], an “average” editor on the English Wikipedia is a young white Anglophone male who is technically inclined, formally educated from a primarily Christian or secular country located in the Global North. Also, based on statistic obtained in 2011 [32], only about 13% of all editors are female, while more than 84% of all editors are between 18 to 30 years old. Therefore, some cultures, topics and perspectives tend to be underrepresented on Wikipedia.

Flöck et al. [22] also provided evidence of bias and lack of methods supporting neutral point of view based on the analysis of previous literature. For instance, they suggest that the longevity of a text fragment can be used as a notion of quality for the edit that introduced it. This can be extended to assess the trustworthiness of the editors, by aggregating the quality of their edits as mentioned before [1, 12, 42, 83]. However, such notions of quality is in favour of old, possibly biased content which can be supported by a majority group confronting a minority group and clearly restricts the replacement of outdated or biased content, especially if to be introduced by new contributors and anonymous editors. The authors also discuss that without appropriate methods to promote diversified and unbiased contributions, new members will get discouraged from contributing to articles and opposing viewpoints will diminish.

Hence, automatic tools and methods that can support producing a balanced and unbiased coverage of topics and information in Wikipedia are needed. Analysis of controversy at different levels is one of the effective approaches towards this goal as reviewed in the next sections.

### **2.3.1 Article-level Analysis**

The research on controversy analysis has been followed by some other authors as well, who attempted to come up with an accurate controversy model, identifying or ranking controversial articles. Kittur et al. [40] were among the first to work in this direction. They used a regression model based on several article-level features to predict the number of dispute tags assigned to an article, and considered these tags as an indicator of the degree of controversy. Examples of these features are the number of editors, the number of revisions, the number of anonymous edits, etc.

Vuong et al. [76] built a model to assign a controversy score to articles assuming a mutual reinforcing relationship between controversy scores of articles and controversy scores of their editors; they also validated their work on the presence or absence of dispute tags. Their intuition is that there are two scenarios where dispute is more serious: aggressive editors on non-controversial topics, and non-aggressive editors on controversial topics. Dispute between editors was also modeled in terms of the number of words added by one editor and later deleted by another editor. However, this can lead the method to mistakenly assign high controversy scores to articles and editors in case of vandalism. For instance, Vuong et al. attributed most of the dispute in the article “podcast” to the conflict between two specific editors, while as observed by Yasseri et al. [82], one of these editors is an anonymous, vandal editor who edited the article only once, but due to his mass blanking action, large dispute values were considered for this editor and another editor fought with this vandalism.

Brandes et al. [8] studied Wikipedia articles from the perspective of editors interactions through a graph-based representation. Interactions of editors are represented by positive or negative edges and are extracted by means of simple intuitive methods such as delete, revert and restore actions. The authors also developed a continuous projection method that partitions editors into two opposing camps. In this way, most of negative edges fall between the two groups rather than within them. Using this partitioning technique, they also proposed a score, called “bipolarity”, that quantifies the degree of controversy of each article by measuring how much partitioning of each graph is close to a bipartite graph having perfect division of editors into the two opposing camps. The results of experiments by Brandes et al. showed average higher bipolarity value for a set of 60 random controversial articles compared to featured articles. However, the authors discussed that the average bipolarity was quite high for both groups of articles, which as examined in more details in Chapter 4 shows that this score cannot be used alone to distinguish controversial articles from other articles.

Yasseri et al. [82] proposed a numeric score based on mutual reverts (when both editors have reverted the work of one another) to model edit-wars showing higher

scores for controversial articles. They also showed examples of consistency of the evolution of the score with external events about a topic, such as the death of the pop singer Michael Jackson in 2009.

Li et al. [45] verified the source of controversy by testing the following hypotheses: 1) The article is controversial because it deals with some specific controversial issues such as “child abuse” and “drug” in Michael Jackson article, 2) The article belongs to a category of inherently controversial topics such as “nuclear technology”, and hence unlike case 1) the whole topic of the article is controversial, and finally 3) the article is controversial due to the aggressive and conflicting behavior of some contributors more than because of the actual content. The authors discussed that if case 2) holds, then we should be able to find topically related groups of controversial articles, and if case 3) holds, then we should expect to see some common contributors across controversial articles. Hence, in this later case, grouping articles based on common contributors should give meaningful categorization. The results showed that none of the cases 2) and 3) can be supported and hence case 1) is more plausible. However, it should be noted that these results are bounded to a small corpus of 68 controversial articles from the “religious” category and further experiments are needed to justify these results.

### **2.3.2 Fine-grained Analysis**

Current controversy models only inform readers and editors about the global state of articles by analyzing articles at the whole article-level. However, what is more important than being able to identify controversial articles is the ability to allow readers to get a general idea of the underlying discussions and debates. Except two very recent works [6, 10], none of the previous methods address this issue.

In both of these works, *wiki links* (i.e. hyperlinks to other Wikipedia’s articles) are used as the unit of analysis of controversy, considering these links to be the set of discussed topics in each article. Moreover, in both of these works, controversy of these units are calculated based on the number of times they have been edited, assuming the highly edited units to be the most controversial units within each article. However, as we discuss in Chapter 4, at least at the article-level, simple heuristics

such as the number of edits or mutual reverts are not sufficient for modeling controversy of Wikipedia articles. In fact, Bykau et al. [10] used the set of “Lamest Edit Wars in Wikipedia”<sup>2</sup> as their validation dataset, which as mentioned by Dori-Hacohen et al. [18] is problematic. This is because these articles contain only those having “edit-wars” and there is a small overlap between this list of articles and list of known controversial articles<sup>3</sup>. Borra et al. [6] also mostly focused on providing different useful user-interfaces for highlighting the most controversial topics within each article, without providing any quantitative evaluation on the correctness of the extracted topics.

In contrast, our methods for fine-grained analysis of controversy not only includes finding the most controversial topics within each controversial article, but also contains finding specific revisions contributing to these controversies. Moreover, we take a more systematic approach for our analyses by formulating both of our methods as a general maximization problem and analyze their computational complexities. We also use the standard list of tagged controversial articles of Wikipedia as our ground truth.

### 2.3.3 Evaluation of Proposed Methods

The proposed controversy models discussed rely on different assumptions and have used different and sometimes limited sample of articles in their experiments, which make it difficult to objectively compare them. For instance, some authors have used the number of dispute tags as their ground truth [40, 76]. However, it is discussed in recent works [66, 82] that dispute tags are not reliable measures due to chance of not adding or removing them right before or after of formation of controversy. Also, different editors might have different tendency in using them, where a community of editors in an article might see a sequence of edits controversial enough to need a tag to be added, while another community in another article might have a higher tolerance for this. Hence, in Chapter 4 we study and compare several different previous controversy models under a standard evaluation scheme and show the

---

<sup>2</sup>[http://en.wikipedia.org/wiki/Wikipedia:Lamest\\_edit\\_wars](http://en.wikipedia.org/wiki/Wikipedia:Lamest_edit_wars)

<sup>3</sup>[http://en.wikipedia.org/wiki/List\\_of\\_controversial\\_articles](http://en.wikipedia.org/wiki/List_of_controversial_articles)

inefficiency of some of them.

## 2.4 Opposing Views in Other Social Media

Another area of related work concerns extracting and/or summarizing opposing views or opinions, which are classical problems in natural language processing [58]. In the realm of social media, however, most of the work is focused on product or movie reviews [20, 58], which are often brief and much simpler compared to Wikipedia articles. Also, most of these methods are supervised, requiring annotated data which would be extremely difficult to obtain in the setting we work on. It is only recently that a few attempts have been made to use unsupervised methods to analyze opinions from more rich text (news articles, in particular), containing more elaborate and subtly expressed opinions compared to review data as explained.

### 2.4.1 Supervised Methods

Many supervised approaches have been proposed for classifying documents into one of the two opposing camps using annotated dispute corpora. Classification of documents can be done at sentence-level [48], document-level [41, 48, 71], or corpus-level [47]. Prototypical example applications of these methods are: posts written by Israeli and Palestinian authors on topics related to the Israel-Palestine conflicts [47, 48], debates related to Bush-Kerry presidential election [47], and on-line ideological debate fora [71]. Despite using different classification algorithms and features (either using all words or focusing on specific lexicons of argument and opinion words), the overall conclusion of these works is that a *viewpoint* is reflected in the choice and in the usage of words. For instance, a viewpoint supporting abortion tends to use words such as “choice”, “women”, and “right” more frequently compared to an anti-abortion view where “baby”, “human”, and “bible” are more frequent words [41]. However, these methods do not apply to our problem directly, as they fail to take the prolonged nature of the dispute, which is a necessary condition for controversy to emerge.

Some supervised methods have also taken into account the social setting sur-



rounding the disputes. For instance, Thomas et al. [74] considered the problem of identifying agreement and disagreement between speech segments in transcripts of congressional debates and meetings. They use many features: textual features, and domain-specific features such as the duration of the speech, the number of speakers between the speech of two speakers and direct mentions to speakers' by name [23]. Supervised methods have also been applied to Twitter data to detect controversial events [60] or assess the credibility of related Tweets [11]. Similarly, these approaches rely primarily on domain-dependent features that are specific to Twitter. For instance, using sentiment lexicons, information about the number of followers of a poster user, the number of times a message gets retweeted, and content of hashtags are some of the most important features used in these works.

Finally, detecting web pages covering controversial issues was studied in a recent work by Dori-Hacohen et al. [17]. In that work, the authors classified each web page to controversial or non-controversial according to controversy labels of top-k nearest Wikipedia articles that are found to be related to it. The controversy labels of Wikipedia articles come from a manually annotated dataset of a small set of articles, and assuming a predefined controversy label for articles that do not have annotated label. The authors also tried using previously proposed controversy metrics such as "Mutual Reverts" [82] in place of manually annotated labels with no success as they found out that their annotated labels for Wikipedia articles did not line up with these automatic scores. These results confirm our hypothesis that covering a controversial topic does not necessarily make an article to become controversial, and the edit process, collaboration and conflict management strategies are other factors affecting article's controversy and hence are important to be considered in Wikipedia controversy models.

## **2.4.2 Unsupervised Methods**

A few recent attempts have been made on extracting opposing views in an unsupervised way from political domains [3, 20, 59]. The common idea in these works is to categorize opinions according to well-known *opinion holders* such as news agencies, political parties, or famous political figures. Note that this is different than

categorizing them into positive, negative relative to each other. Their argument is that, in political texts, opinions are expressed in a much more complex form compared to evaluative and review texts. In fact, in political text, often it is the choice of words and the arguments that differentiate two opposing views instead of having positive or negative attitudes towards the same issues. For instance, the two statements of “we want responsible healthcare reform based on private insurance“, and “we want universal healthcare reform with a public government-run health insurance agency” stated by Republicans and Democrats respectively both can be viewed as a general positive opinion. However, there is a huge difference between the viewpoints of these two parties on this issue reflected in words such as “private” and “responsible” vs. “universal” and “public” [20].

Fang et al. [20] proposed a perspective-based topic model to extract the common topic words (i.e. words related to the background topic and not depend on any perspective) and opinion words (i.e. words indicative of a specific perspective) across a set of text collections coming from different perspective sources. The perspective in this work is assumed to be known and is modeled as a dominant opinion group such as “democrat” vs. “republican”, or “New York Times” vs. “The Hindu” (i.e. a news agency in India). Given the set of documents of each perspective, corresponding viewpoints were represented by a set of opinion words, and the difference between different perspectives across various topics were quantified using a proposed diversity metric. For instance, two perspectives might be found to have more similar opinions on “abortion” compared to “immigration” issue. Also, issues having diverse viewpoints across different perspectives, which is measured by the diversity metric, were considered to be controversial.

Park et al. [59] suggested viewing controversial topics from the “opponent-based” view. In this work, the idea is to first identify the two opposing groups and main opinion holders for each given topic in a news document and use this information to classify the given news document into one of these two groups. The opinion holders in this work are extracted from the subject of direct or indirect quote statements. However, topics are *assumed* to be contentious and a set of news documents related to each topic are assumed to be available. Hence, the task is to

just assign each of these topic-related documents to one of the opposing camps or “none” if a specific side is not supported.

A similar opponent-based view can also be seen in [3], where the aim is to build a network of opinion holders, and their sentiments towards public political events. In that work, opinion holders are named entities appearing as the subject of opinionated text snippets. Opinionated statements are found initially by a set of seed opinion patterns such as “he supports”, “he opposes”, and later on expanded by using a diverse set of patterns, found iteratively from the previously found opinionated statements. The acquired opinionated statements are canonicalized and organized in a hierarchical order of topics to form a network of opinion holders and opinion targets (the set of words explaining the context of the opinion) connected by a positive or negative attitude. For instance, topic of “Conflict in Syria” has been categorized into several subtopics such as “arming Asad’s regime” and “arming rebels in Syria”, and opinions of public figures such as Barack Obama, Hillary Clinton, and Russia with respect to these topics in the form of “support” or “oppose” were extracted.

There are many fundamental differences between our work and the methods discussed above. First, those methods did not try to classify topics as controversial, and focused only on extracting opposing views out of a set of documents related to a topic. Moreover, the extracted opposing views are only from the point of view of known political figures or parties, which is different from the opposing views that can be extracted from Wikipedia. In fact, by focusing only on the political domain, these works assumed not only that there are opposing viewpoints but also that they are sufficiently divergent to justify analysis. In reality, this assumption might not hold for most controversial topics. Moreover, the evolutionary nature of Wikipedia articles is completely different from the open, and diverse language used in news and arbitrary web documents. This diversity in language and word usage are the main factors used in previous studies to identify opposing views and perspectives as explained earlier [20, 41, 48, 71].

## Chapter 3

# Identifying Controversial Articles Using Collaboration Networks

Some authors looked at some of the issues that arise during the lifecycle of typical Wikipedia articles due to differences of opinion among editors. Flöck et al. [22] discuss several problems such as resistance against new content from “occasional” editors, the difficulty in changing the content in stable and mature articles, and cases with strong feeling of ownership and defensive behaviour of some editors. They argued that such issues have a negative impact on the diversity and NPOV in Wikipedia. Brandes et al. [7] studied some of the factors that lead to editors to stop contributing to their articles, and showed that editors of controversial articles are more likely to quit Wikipedia. One explanation for this phenomenon is the frustration of being involved in long debates, vandalism and edit-wars.

The problems mentioned above emphasize the importance of mechanisms to help editors and readers to detect and understand the differences of opinion that lead to controversy. Manual tagging of controversial articles, clearly, is not an ideal solution. What is needed are consistent and scalable methods that can be deployed automatically. Towards this goal, in this chapter, we describe an effective method for identifying controversial articles and distinguishing them from other articles.

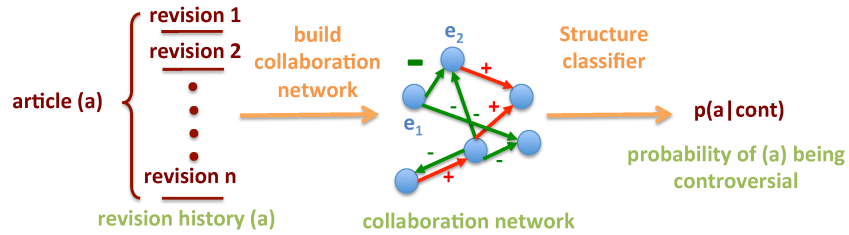


Figure 3.1: Overview of our proposed method for identifying controversial articles

## 3.1 Proposed Method

### 3.1.1 Overview

Our approach for modeling controversy and identifying controversial articles in Wikipedia is based on a user-driven method focusing on the set of editors of each article and the type of collaboration relationships among them. Figure 3.1 illustrates the main steps in our approach. Given an article  $a$ , we start by building an internal structure referred to as *collaboration network*. The network consists of all editors of the article (vertices) and the pairwise *attitude* of the editors that have a collaboration on the article. Next, we map the collaboration network of article  $a$  to a notion of controversy for that article using a classifier called “Structure classifier”, which is trained on a sample of real Wikipedia articles. The output of the classifier on unseen articles is a probabilistic assignment of the article to class of controversial articles. If this probability is higher than a threshold, the given article is considered to be controversial, and non-controversial otherwise.

In the next sections, we first define collaboration networks more precisely, and then present the details of the Structure classifier.

### 3.1.2 Collaboration Network

**Definition 1** A collaboration network is a directed, signed graph  $G = (V, E)$  associated with a Wikipedia article  $a$ , where  $V$  is the set of contributing editors and  $E \subset V \times V \times W$  is the set of weighted edges connecting editors whom there exists a collaboration relation between them.

This definition of collaboration network is general and different types of net-

works can be built based on it. For instance, the set of contributing editors can include all editors of the article, or it can be limited to only those who have edited at least a specific number of revisions. Many other ways depending on the definition of *contribution* are possible. Similarly, there are different possibilities for defining collaboration relation between editors. For instance, we might consider editor  $e_1$  to have a collaboration relation with editor  $e_2$ , if  $e_1$  revised a revision of  $e_2$  in some way. Alternatively, this relation can be defined based on the time difference between edited revisions.

Regardless of how collaboration relation is defined, we consider this relation to represent attitudes of editors towards one another. More formally, collaboration relation is a function  $R(e_1, e_2) \rightarrow [-1, 1]$  that assigns a real number in the given scale, where positive numbers indicate agreement and supportive attitude, while negative numbers show disagreement and opposing attitude. Zero value indicates that there is no collaboration relation between the given editors, or the type of this relation is unknown for us.

In Section 5, we present one approach for building collaboration networks, which includes a novel and effective way for inferring attitudes of editors.

### 3.1.3 Structure Classifier

Once we build the collaboration networks of the articles, we need to map them into a measure of controversy. Intuitively, we want this mapping to give on average higher controversy levels for articles identified as controversial by human editors than non-controversial articles, providing a clear separation of the two kinds of articles. Hence, we need to look for a property or a combination of properties that can reflect the main differences between the overall structure of networks of controversial and non-controversial articles.

Previously, Brandes et al. [8] showed structural difference of controversial and non-controversial articles using a measure called *bipolarity*. Bipolarity is a graph-based measure that indicates how likely it is to decompose a graph into two partitions representing opposing groups, where most of negative edges will lie across the partitions rather than within them. Intuitively, the collaboration network of a

controversial article would approximate a perfectly bipartite graph: each partition would correspond to an opposing group of editors holding an opposing view compared to the other. Hence, one approach would be to compute bipolarity from the collaboration networks we build and compare its values over controversial articles and non-controversial articles. However, bipolarity is defined only for negative edge networks, and was also shown to not provide enough discrimination between controversial and non-controversial articles [66]. Therefore, we extracted some other features from our collaboration networks instead of focusing on a single metric.

In particular, we rely on structural properties of networks of controversial and non-controversial articles. These properties provide insight about distribution of nodes and edges (positive and negative) and how in general each network looks like. We use these properties as features to train the Structure classifier, and identify controversial articles using this classifier. Our features include several metrics derived from social network theories of collaboration, which clearly help in the prediction accuracy of our method.

### 3.1.3.1 Structural Features

We extract the following features from the collaboration network associated with each controversial or non-controversial article:

- total number of nodes (*nodes*)
- total number of (*edges*), positive (*edges*<sup>+</sup>), and negative(*edges*<sup>-</sup>) edges
- average of total (*avg\_degree*) degree of nodes
- average of positive (*avg\_degree*<sup>+</sup>) degree of nodes  
item average of negative (*avg\_degree*<sup>-</sup>) degree of nodes
- fraction of nodes whose degree is higher than  $0.9H$ , where  $H$  is the highest degree in the network, for each of the categories; positive in-degree (*high\_in*<sup>+</sup>), negative in-degree (*high\_in*<sup>-</sup>), positive out-degree (*high\_out*<sup>+</sup>) and negative out-degree (*high\_out*<sup>-</sup>)

- fraction of nodes whose degree  $d$  satisfies  $0.9M \leq d \leq 1.1M$ , where  $M$  is the mean of the node degrees in the network, for all categories as above:  $mid\_in^+$ ,  $mid\_in^-$ ,  $mid\_out^+$  and  $mid\_out^-$
- fraction of nodes whose degree is less than  $1.1L$ , where  $L$  is the lowest degree in the network, also defined for all categories:  $low\_in^+$ ,  $low\_out^+$ ,  $low\_in^-$ , and  $low\_out^-$
- fraction of nodes with more positive than negative incoming edges ( $more\_in^+$ ) and outgoing edges ( $more\_out^+$ )
- total number of triads ( $triads$ ) in the network
- the relative number of each of the 8 triad types ( $triad_1, triad_2, \dots, triad_8$ )

The features concerning the in and out degrees of positive and negative edges are meant to reflect the skew in the distribution of these kinds of edges. For example, a collaboration network in which only a small fraction of editors has the majority of the negative edges would be a potential sign of controversy: those editors consistently disagree with the others, and such disagreement is reciprocated. The opposite situation would be a network in which there is no skew in these distributions, possibly indicating that eventual disagreements are the result of reasonable differences of opinion.

Triads are directed sub-graphs of size 3, which have been used as important metrics in social network analysis [21,24,43]. For instance, “balance theory”, based on the common principles that “friend of my friend is my friend” and “enemy of my friend is my enemy” is one of most known examples of social-psychological theories linked to triads distributions in real social networks [43].

In our work, we considered eight different triad types, shown in Figure 3.2 depending on how many negative edges exist (0, 1, 2, or 3), and whether the edges in the triad form a cycle or not. Triads 1,3,5 and 7 are plausible according to the “balance theory”.



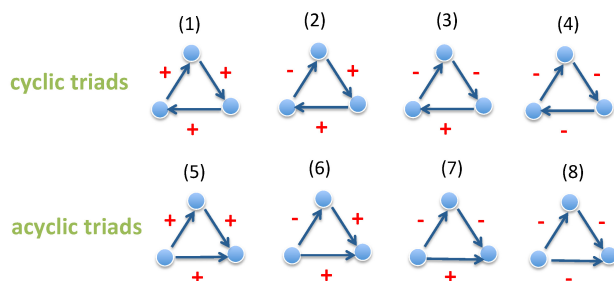


Figure 3.2: Eight different triad types used in our Structure classifier

## 3.2 Experimental Results

In this part, we report the results of our experiments using the Structure classifier in the task of identifying controversial articles.

### 3.2.1 Dataset

We selected 240 articles for each class of controversial and non-controversial, and extracted their entire revision history from the Wikipedia dump dated at March, 2011<sup>1</sup>. We chose these articles as follows.

For the controversial category, we selected articles randomly from the list of articles manually identified as controversial by the Wikipedia community<sup>2</sup>. The chosen articles account for 1/3 of all articles listed as controversial at the time the data was collected. These articles are selected randomly from all different 15 categories of topics in a way to have roughly the same number of articles from each category. Examples of these categories are History, Religion, Science, Philosophy, Sport, etc.

In this way, our dataset is representative of different controversial topics that can be found in Wikipedia and is based on a gold standard that contains all articles with known controversial issues, and not only those which have dispute tags. This is because while it is reasonable to assume that an article with many tags is controversial, a low or zero value does not necessarily mean lack of controversy [82]. In fact, out of the 240 articles we selected from the above list, only 122 articles had dispute tags at some point in their revision history. Moreover, there are other issues

<sup>1</sup><http://dumps.wikimedia.org/enwiki>

<sup>2</sup>[http://en.wikipedia.org/wiki/List\\_of\\_controversial\\_articles](http://en.wikipedia.org/wiki/List_of_controversial_articles)

with judging the degree of controversy of articles based on the number of dispute tags, as discussed in previous work [66].

For the non-controversial category, we picked (randomly) 100 articles from the featured category and 140 articles from other quality categories. We checked and discarded any article that had been tagged as controversial at any point in their edit history. This is because many of the articles in the list of controversial articles later become non-controversial, , and some even improve to featured articles.

### 3.2.2 Comparison with other methods

The Structure Classifier we described is general and can be applied to any type of collaboration networks. We apply it on networks built using different methods and compare it with a baseline method, and a method that uses statistics extracted from revision history of articles. More specifically, we study the following methods in this chapter:

1. **DRR:** In this method, we apply the Structure classifier on DRR (Delete, Revert, Restore) networks. These networks are based on the work of Brandes et al. [8], who used a notion of collaboration networks in their work that considers a node for each editor who has contributed at least one revision. Also, they used delete, revert and restore actions for assigning signs to the edges. More specifically, in their work, whenever editor  $e_1$  deletes some words originally inserted by  $e_2$  in the text of article, an edge with a negative weight proportional to the number of words deleted is created from  $e_1$  to  $e_2$ . Also, whenever  $e_1$  restores a version created by  $e_2$  to an earlier version created by  $e_3$  (a possibly different editor than  $e_1$ ), a single unit positive edge is created from  $e_1$  to  $e_3$ , and a single unit negative edge is created from  $e_1$  to  $e_2$ . This is because  $e_1$  had undone the work of  $e_2$  and implicitly agreed with the work of  $e_3$ .
2. **PV:** In this method, we apply the Structure classifier on PV (Profiles and Votes) networks. PV networks are built by inferring the attitudes of editors by employing a classifier that uses history of collaborations of editors, signs

of votes (support or oppose) in admin elections. The detail of this method is explained in more details in Section 5.

3. **Rand50:** In this method, we apply the Structure classifier on randomly generated networks. For generating these networks, we use the same structure as in PV networks, but the signs of edges are assigned randomly. In this way, we can see how the Structure classifier performs when it lacks the type of relations (i.e. positive or negative) between editors, and only has access to the structure of the network of their collaborations.
4. **NE-count:** Rand50 has access to entire structure of the collaboration networks, and only lack the type of relations of editors. In NE-count, we limit this information further by only using the number of nodes and the number of edges. We extract these two features from the networks built according to PV method.
5. **Meta classifier:** The Meta classifier [65] is a classification-based method that uses a set of features extracted from the revision history of articles. These features are as follows (AVG, STD, MAX mean average, standard deviation and maximum respectively): absolute number of (1) revisions, (2) minor revisions, (3) revisions by anonymous editors, (4) unique editors, (5) anonymous editors; percentage of (6) anonymous editors, (7) revisions by anonymous editors; (8) ratio of number of revisions to unique editors; AVG, STD and MAX for the (9) number of revisions per editor; and AVG, STD of (10) length of edits in each revision. Also, there were more complex features based on considering a type of disagreement relations between editors. We excluded these features to only focus on meta features in this chapter.

### 3.2.3 Classification Accuracy

Table 3.1 shows the accuracy of the methods we studied in detecting controversial articles. The results are based on 10-fold cross validation using Logistic classifier in Weka <sup>3</sup>. The same 30 features were used in DRR, Rand50, and PV, and the

---

<sup>3</sup>[www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)

Table 3.1: Results of identifying controversial articles

Method	Accuracy
NE-count	56.70%
Meta classifier	75.20%
DRR	64.31%
Rand50	68.67%
PV	<b>84.58%</b>
PV + Meta classifier	<b>89.12%</b>

difference between these methods is only in the different networks that each method uses. Also, the results for Rand50 are the average results over 20 runs with different random seeds.

First, as we can see the NE-count baseline has the lowest accuracy among all methods showing that the number of editors and their collaboration relations are not enough to distinguish controversial articles from other articles and the actual network structure matters.

Second, we see that the Structural classifier when applied on some types of collaboration networks can result in very high accuracy. In particular, the combination of the Structure classifier and PV networks produces the highest results among the studied methods. Comparing this method with Rand50, where the same network structure is used, and just the sign of edges is different across these methods is indicative of the important role of correctly inferring attitudes of editors. Also, the significant difference of PV method and DRR method, where the structures of networks are also different, shows the effectiveness of PV method for building collaboration networks. More specifically, DRR includes all editors of an article into its collaboration network, and considers collaboration relations of editors at the word level and based on basic edit operations. In contrast, PV excludes occasional editors from the network of each article, and uses an extensive set of global features to *learn* to infer the type of collaboration relations between editors as described in Section 5.

Finally, comparing with the Meta classifier, we see that while general features

about the revision history provide a good discrimination between the two studied classes of articles, they cannot eliminate the important role of the structural properties of the collaboration network of editors. In fact, by taking advantage of these two complementary views (structural and meta features), we are able to boost the performance of both methods further and achieve a very promising results of 89.12%.

In the rest of this chapter and the next chapter, we study the Structure classifier applied on PV networks in more details as this type of collaboration networks resulted in the highest accuracy among all types of collaboration networks we studied. For simplicity, if not explicitly stated otherwise, the Structure classifier method refers to applying this classifier on PV networks.

### 3.2.4 Feature Analysis

In this part, we study the effect of different features on the performance of the Structure classifier. Our model for learning this classifier was based on Logistic Regression, which learns the relationship between the probability of a dependent variable (the modeled class) and one or more independent variables (features) as follows:

$$p(y|f) = \frac{1}{1 + e^{-(b_0 + \sum_{i=1}^n b_i f_i)}} \quad (3.1)$$

where  $y$  is the dependent variable,  $f$  is a vector of features  $(f_1, f_2, \dots, f_n)$ , and  $b_0, b_1, \dots, b_n$  are coefficients associated with each feature learned from training data. More specifically, each  $b_i$  shows the change (increase when  $b_i > 0$ , decrease when  $b_i < 0$ ) in the log odds of occurrence of the modeled category (i.e. log of ratio of probability of the modeled category to probability of the other category) for a one-unit change in  $f_i$ . Hence, to study the effect of each feature on our modeled event (having a controversial article), we used  $b_i$  coefficients after standardizing all feature vectors to have the same scale for different features.

**Results** Figure 3.3 shows a comparison of these coefficients for the top-10 features, where larger coefficients correspond to having more effect on changing (increasing or decreasing depending on the sign of the coefficient) the odds of having

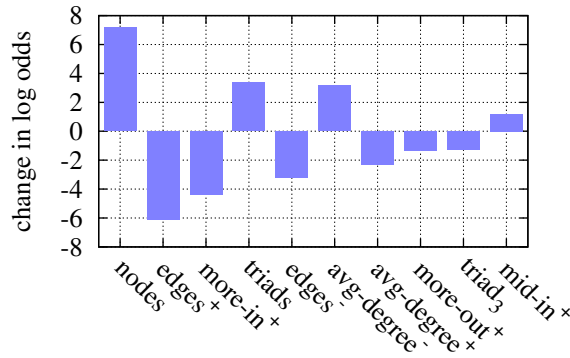


Figure 3.3: Effect of top-10 features (positive values on y-axis increase the odds of controversial class, while negative values decrease it)

a controversial article.

As we can see, *nodes* is the feature with the highest contribution and it has a positive effect on increasing the odds of controversial class. This is expected, since as the number of editors increases, the diversity of points of view and perspectives on the topic increases as well, and thus there will be more chance of conflicting opinions to be brought forward.

Other features to note are *edges<sup>+</sup>* and *edges<sup>-</sup>* where they both have negative effect on odds of the controversial class. At first glance, we might expect to have different effects for these features with the intuition of having more positive edges in non-controversial articles than controversial articles and vice versa. However, it should be noted that a network of a controversial article is expected to consist of positive (agreement) edges within editors who agree with each other, and negative (disagreement) edges across editors who disagree with each other. It is the *structure* of the network and formation of these edges that determine whether the article is controversial or not. Another point to consider is that the number of positive and negative edges is not necessarily indicative of the number of times that two editors interacted with each other, as multiple interactions between the same pair of editors are combined and represented as one single positive or negative edge in our work. Hence, while we expect to have more total interactions in controversial articles due to their average longer revision history, having fewer edges does not imply having less interactions between editors as explained.

Interestingly, the difference of positive and negative edges across controversial and non-controversial articles is reflected in the difference of average positive and negative degrees of nodes. As we see  $avg\_degree^+$  has a negative effect on controversial class, while  $avg\_degree^-$  has a positive effect on this class. This suggests that controversial articles are more likely to have editors with higher negative degree, and lower positive degree than non-controversial articles. Several other degree-related features such as  $more\_in^+$  and  $more\_out^+$  are also among the top-contributing features showing the importance social network analysis for this task.

Finally, we see that the total number of triads is one of the most contributing features for our classifier, which corroborates the hypothesis that understanding the collaborative editing of Wikipedia (which is by definition a social process) is key to detecting controversy. Among the different triad types,  $triad3$  is the most significant feature. This triad type corresponds to a setting where all edges are positive and form a cycle, which is consistent with the common principle of “a friend of my friend is also my friend”. As the coefficient associated with this triad is negative, it seems that this principle is observed more in networks of non-controversial articles. That is, in non-controversial articles we are more likely to observe editors “helping” each other promote their revisions and the point of view they want to convey in a perhaps more constructive or less combative way.

### 3.2.5 Effect of History Length

We also consider the effect of the length of the article’s history on the accuracy of our method, with two experiments. For clarity, we focus on the two best performing methods, the Meta and Structure classifiers. In the first experiment, we divide the articles into three bins, based on the length of their revision history: the top 1/3 articles with the *longest* histories; the bottom 1/3 with the *shortest* histories; and the ones left, with *average* histories. We then proceed with the usual 10-fold cross validation as before, but report the accuracy for each group separately in Table 3.2. For the second experiment, we artificially cut-off the revision history of each article after different points, and assess the effectiveness of the method with this limited history (Figure 3.4).

Table 3.2: Accuracy results of each method on different ranges of the edit history

	shortest histories		average		longest histories	
	contr.	non-contr.	contr.	non-contr.	contr.	non-contr.
Meta	0.65	0.88	0.81	0.73	0.79	0.52
Structure	0.85	0.87	0.82	0.85	0.94	0.75

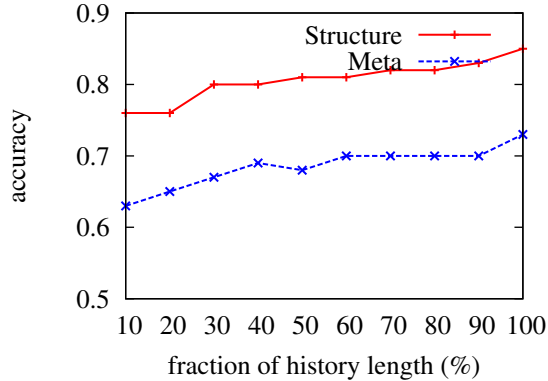


Figure 3.4: Effect of limiting history length

**Results** Table 3.2 shows the results of the first experiment, where we can see that our method maintains a relatively high accuracy for both classes of articles across all edit ranges. In comparison, the Meta classifier has quite poor performance on low-edited controversial, and high-edited non-controversial articles. This is not surprising as controversy is generally correlated with higher number of revisions (i.e. most controversial articles have high number of revisions, and most non-controversial articles have low number of revisions) and detecting articles that are different from this general trend is more difficult for a classifier that uses this feature. On the other hand, the Structure classifier has much higher accuracy on these two categories of articles, especially on low-edited controversial articles where it outperforms the Meta classifier by about 20%. While generally the size of the networks (in terms of the number of nodes and edges) is smaller for these low-edited articles, these results suggest that our method is not affected by this factor very much and still is able to have high accuracy in this range as well.

The results of the second experiment are shown in Figure 3.4, where each point corresponds to one of the history periods explained before. As can be seen while increasing the length of the history overall has positive effect on the accuracy of our



method, it is worthwhile to note that our method has a reasonably high accuracy even at the smallest history periods. In particular, at 10%, our method has 75.67% accuracy which is about 14% higher than the Meta classifier. This shows that not only the Structure classifier outperforms the Meta classifier on the entire revision history, but it has the capability of working significantly better when using only a small portion of the history of each sample article.

Overall, the results of these two experiments show that the Structure classifier is very effective, even for articles with short histories or when applied only on a small fraction of the article’s entire revision history. This is significant, as it indicates our method could be very effective on samples of the edit histories of the articles. Moreover, these results indicate our method is not biased towards longer histories, which is, alone, a useful baseline to predict controversy. These results are evidence of the power of our collaboration networks in capturing the editorial process in Wikipedia. They also make evident the benefits of extracting features rooted in sound social theories instead of relying on simple heuristics and features that are just easy to extract, as done by many previous methods.

### **3.2.6 Filtering Networks by Active Editors**

The final experiment we report studied the effect of filtering the collaboration networks by removing less active editors. For this purpose, we ranked contributed editors of the collaboration network of each article in terms of their number of contributed revisions. We built several networks, keeping only a certain fraction of the top editors according to that ranking.

Figure 3.5 shows the accuracy of the method for the different sub-networks. As expected, the accuracy increases as we increase the fraction of editors in each network. Also, the relative increase in accuracy seems to taper off over time. To achieve 0.75 accuracy<sup>4</sup>, one needs to use only the 30% most active editors to build the collaboration networks. This indicates that the editors that are prone to cause controversy tend to be fairly active as well.

---

<sup>4</sup>Note that this accuracy level is higher than what any previous method achieved in our tests.

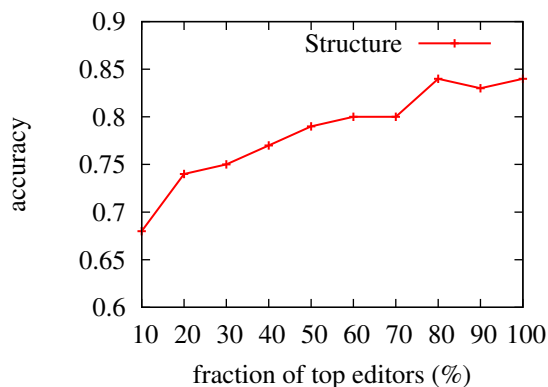


Figure 3.5: Effect of filtering collaboration networks by top active editors

### 3.3 Conclusion

In this chapter, we detailed our method of modeling controversy of articles in Wikipedia, which relies on analyzing editors collaborations in the form of collaboration networks. Our controversy model was based on extracting structural features from collaboration networks of articles and training the Structural classifier.

We applied this classifier on different types of collaboration networks using different structures and methods to infer the signs of edges. Among these different network types, we obtained the best results using PV networks. As explained in Section 5, in these networks, an effective method for learning to infer the attitude of editors has been employed. This when combined with structural features used in the Structural classifier, enabled us to have a highly accurate and effective method for identifying controversial articles as shown by several experimental results and comparison with other network-based and classification methods.

These results also showed that the Structural classifier is successful when limiting the size of revision history that is used to detect controversy of an article, or when working with articles having short history. This behavior can be attributed to the internal network structure that our controversy model uses which is in contrast with directly applying features about each article as done for instance in the Meta classifier. We also found that controversial articles in our method can be identified using only the most active editors of the collaboration network of each article.

## Chapter 4

# Comparative Study of Other Controversy Models

There have been a few recent attempts to address the problem of identifying controversial articles in Wikipedia [8, 40, 68, 76, 82]. Most of these methods aim at providing a single controversy score which is then used in classifying or ranking articles. However, various methods were evaluated using different criteria and on different sets of articles by different authors, making it hard for anyone to verify the efficacy and/or compare different methods. For instance, Brandes et al. [8] studied only 60 random controversial articles, while Kittur et al. [40], and Vuong et al. [76] focused only on articles about religion. Sumi et al. [82] used a simplistic model of controversy, concluding that the complexity of detecting controversy in Wikipedia has been over-estimated and there is no need for designing complex models. However, they neither used a standard evaluation strategy, nor did they compare their results with previously proposed methods such as the work of Kittur et al. [40] and Vuong et al. [76].

In this chapter, we attempt to close the gaps indicated above. We study and compare different models of controversy under a standard framework and in terms of different metrics. In particular, we show that while some methods are simple and intuitive, in practice the underlying process of controversy formation in Wikipedia articles is too complex to be captured by these heuristics. Thereby, identifying controversial articles out of a pool of non-controversial articles needs to employ more sophisticated methods such as machine learning tools, where controversy is

detected by using a combination of factors learned from some annotated examples. Our Structure classifier introduced in Chapter 3 is an example of one of these supervised machine learning methods that we examine it in more detail in this chapter by comparing it against previous controversy models. We also discuss a categorization of different models in terms of resources and techniques they use to provide a perspective on designing future, improved controversy models.

## 4.1 Examined Methods

We now discuss the five methods we compare. What is common in all of these methods is that they all rely on simple numeric features extracted from the revision history of the article or its discussion page without analyzing the textual content of the pages.

Table 4.1 summarizes the main characteristics of the studied methods in terms of the model used for disagreement and controversy. The following sub-sections give more detailed description of each of these methods.

Table 4.1: Summary of the main characteristics of the studied methods

Method	Disagreement model	Controversy model
Mutual Reverts	mutual reverts	mutual reverts
Bipolarity	deletes + reverts	closeness to a bipartite graph
Basic REA	deletes	ratio of deletes to all contributions
Structure classifier	inferred attitudes	statistics from collaboration networks
Meta classifier	-	statistics from article and discussion page

### 4.1.1 Mutual Reverts

Mutual Reverts(MR) is a single score intended to quantify and rank the degree of controversy of Wikipedia articles [82]. This score relies on revert actions as the direct sign of disagreement and dispute between editors. As reverts are also common in combating vandalism in non-controversial articles, the authors focused only on *mutual reverts*, where two editors have reverted each other’s edits at least once. To account for different activity rates of different editors, and to filter out less active

editors such as vandals, the method considers the minimum number of edited revisions of each editor in each pair of mutually reverting editors. In this way, disputes among “occasional” editors such as vandals get less weight than those involving regular and typically more passionate editors. Moreover, the method avoids “personal” conflicts restricted to two specific editors by ignoring the maximum conflict score (of all pairs) within each article. Finally, the total number of distinct editors engaged in mutual reverts is considered as another important factor in heating the debates. The final score is as shown in the following equation:

$$MR^a = E \times \sum_{N_i^a, N_j^a < max} \min(N_i^a, N_j^a)$$

In this equation,  $MR^a$  refers to MR score of article  $a$ ,  $E$  is the total number of editors, and  $N_i^a$ , and  $N_j^a$  are the number of revisions made by editors  $i$  and  $j$  who mutually reverted each other’s revisions at least once in this article. Also,  $max$  is a constant equal to the largest value of the  $\min(N_i^a, N_j^a)$  across all of these editors to filter out the pair with the maximum conflict score.

This simplistic metric relies on information that is easy to extract: reverts, and the number of revisions of each editor, making it fast and easy to calculate, compared to other metrics we study in this chapter. Also, these simple factors allow the model to work across different Wikipedia languages. In addition, the authors showed that this simple metric outperforms several different single metrics such as the number of authors or the size of the discussion page in ranking controversial articles.

However, for their evaluation, the authors only considered the precision in the top-30 ranked articles returned by scores of each metric. While it is expected that the top scores arise from controversial articles, precision in mid and low ranges of values were not tested. For instance, the authors reported the percentage of controversial articles for different values of scores. For values below 180, 50% of articles are controversial, while this ratio is 60% for the values under 1000, which shows that in both of these ranges, both controversial and non-controversial articles have the same likelihood. Of course, with increasing the threshold of scores, precision

increases, but the recall in the sense of finding examples of controversial articles at the same time decreases. This is why it is important to consider both discrimination ability and performance across both classes of controversial and non controversial articles for the evaluation of such methods.

### **4.1.2 Bipolarity**

Bipolarity [8] is a single numerical score extracted from the “collaboration networks”, built according to DRR method discussed in Chapter 3. As discussed in that chapter, DRR collaboration networks are built by considering the number of words restored from a reverted edit as the weight of positive edges, and a combination of the number of reverted edits and deleted words as negative edge weights.

Once with the DRR collaboration network, the Bipolarity score is calculated by only taking into account negative edges which represent disagreement between editors. The score ranges between 0 and 1 representing how much the collaboration network of an article is close to a fully bipartite graph. The higher the score, the more similar the graph is to a perfect bipartite graph, and the more likely it is to have two opposing camps of editors where most disagreement edges are between the two camps rather than within each.

The authors showed that, on average, controversial articles have higher bipolarity scores than featured articles, which is quite consistent with the intuition that one might have about the formation of controversy where bipolar structure of editors is expected. However, as pointed out in our previous work [65] the variances of bipolarity scores for these two classes of articles is quite high, which limits the applicability of bipolarity for distinguishing controversial articles.

This limited ability can be attributed to the approach taken for building collaboration networks. First, Bipolarity works only with the negative (disagreement) edges and does not take advantage of positive edges that have been shown to be important in some previous works on signed networks [43]. Second, in inferring the weights of disagreement edges, a simple model based on only deleted words and revert actions have been adopted, which can limit the effectiveness of this method compared to more sophisticated models. In particular, as noted by Brandes et al.

both delete and revert actions are seen in featured articles as well. What is worse, combating vandalism is common in featured articles, leading to high bipolarity scores for these articles. Hence, at the very least, these edit actions are not effective to distinguish between dispute-based disagreements, and vandalism-based disagreements.

### 4.1.3 Basic REA

Vuong et al. [76] proposed one of the first methods specifically targeting the problem of controversial articles in Wikipedia. In their work, they described three different controversy models, where two of them are based on a *reinforcing editor-article (REA)* relation. These two models consider a controversy score for editors in addition to considering scores for articles; ; moreover, the two scores are defined in a mutually-reinforcing way. More specifically, they assume that a dispute is more serious when it happens between aggressive and combative editors who have high controversy scores on less controversial articles, or between editors with low controversy scores on articles with high controversy scores. Hence, at each step, the controversy score of an article is updated by the amount of “dispute” that happened between each pair of opposing editors weighted by the controversy scores of these editors at that step. Next, the controversy score of each editor will be updated based on the updated controversy of the article edited by him/her, and this dual updating process continues until convergence. Dispute between each pair of opposing editors is considered as the number of words that were written by one editor and deleted by the other.

Using the same dispute model, the authors also proposed a simpler approach in their paper referred to as *basic* model. This approach is not reinforcing, and controversy is calculated as follows:

$$C_a = \frac{\sum_{i,j} d_{i,j}^a}{\sum_i o_i^a}$$

where  $C_a$  is the controversy score assigned to article  $a$ , and  $d_{i,j}^k$  and  $o_i^k$  are the disagreement values between each pair of editors  $i$  and  $j$ , and the number of words authored by author  $i$  respectively.

Hence, the basic model is just the ratio of deleted words to all contributed words, where it is expected that controversial articles have higher ratios. The authors showed that the reinforcing-based methods have better performance than this basic model. Unfortunately, the computational cost of their reinforcing methods was so high we could not apply them on our dataset even after weeks of computing time. The main reason is that Vuong et al. [76] focused on a specific category of articles, where there is a large number of common articles for each pair of editors which makes to have a small number of articles to be processed for each editor in the reinforcing updating procedure.

However, on our dataset, articles were sampled from very different categories, where the chance of finding common articles between target editors (i.e. editors contributed to test articles) is very low. In order to calculate the score of these target editors while recursively calculating the scores of the target test articles, one has to process a very large bipartite graph of articles and editors. For instance, in the first layer which is where we have target editors, we had examples of editors with thousands of edited articles, where expanding these thousand articles at the second layer can add hundreds of thousands of articles and editors.

It should be noted that aside from the excessive computational demand to process and update scores recursively on this big graph, the convergence of scores can be the another reason of our unsuccessful attempt. In particular, these mutual scores do not follow the general template of HITS-like algorithms that have convergence guarantees. The authors also did not provide any proof of the convergence of their method.

Therefore, with not being able to test the reinforcing-based approaches, we focused on studying the basic proposed model which for simplicity is referred to as “Basic REA” model in the rest of this chapter.

#### **4.1.4 Structure classifier**

The Structure classifier is the method we proposed and described in Chapter 3. For completeness, we briefly review its main details. Similar to Bipolarity, this method also works on collaboration networks. However, compared to Bipolarity, the Struc-



ture classifier makes use of both positive and negative edges in the collaboration network of each article. In addition, in this method, collaboration networks are not represented by a single metric, but rather by extracting the following groups of features from each network.

- basic features such as number of nodes, number of positive edges, etc.
- degree distribution features such as the percentage of nodes having an in-degree of higher than 90% of maximum in-degree, and similarly for out-degree
- triad features including 8 different types

Since, networks are represented by a feature vector, the final controversy model is based on a classification approach where the feature vector representation of collaboration networks are learned to be controversial or not. This classifier can be applied on any signed collaboration networks. However, as tested in Chapter 3, the PV collaboration networks resulted in the best performance of this classifier, and hence we study this classifier using these networks. The main characteristic of these networks is that the signs of edges are inferred using an extensive set of global features representing edit behavior of editors, along with votes cast in admin elections.

#### **4.1.5 Meta classifier**

The meta classifier proposed by Kittur et al. [40] is another classification approach for identifying controversial articles which relies on extracting a set of objective statistics from the revision history of an article or from its discussion page. The authors proposed 30 different features including the number of revisions of an article, the number of unique editors, the number of out-link, and in-links, etc. and found the following seven features as the most important features:

- number of revisions of the discussion page
- number of minor revisions of the discussion page

- number of unique editors of the discussion page
- number of revisions of the article
- number of unique editors of the article
- number of revisions of the discussion page by anonymous editors
- number of revisions of the article by anonymous editors

While none of the other methods considers the discussion page associated to an article in modelling controversy, we can see that more than half of the most important features in this method are related to statistics of such pages. Kittur et al. [40] also emphasized the importance of discussion pages in their paper by showing that there has been a trend of less direct edits on articles, and instead more edits and discussion on discussion pages. However, this trend only was studied on English Wikipedia and as suggested by some other works, discussion pages are less active in other languages [82].

Also, it should be noted that a different meta classifier was later proposed in our work [65] using some of the meta features from Kittur et al. and some new features, where all features were extracted from the revision history of the article itself. This is the classifier used in experiments of Chapter 3. The two meta classifiers are comparable in terms of accuracy with our classifier achieving slightly better results. However, as the work of Kittur et al. [40] is representative of one of the well-known early studies on controversy in Wikipedia, and is it the only method that considers discussion pages prominently, we focus on their classifier in this chapter.

## 4.2 Evaluation Framework

In order to study different methods, we consider the binary classification problem of determining whether or not a specific article is controversial. Some of the methods we evaluate aim at *ranking* articles based on their degree of controversy, where the goal is to predict the number of dispute tags an article should get. However, judging the degree of controversy of articles with these tags is problematic due to

several reasons. First, while it is reasonable to assume that an article with many tags is controversial, a low or zero value does not necessarily mean lack of controversy [82]. For instance, many of the articles on the list of controversial articles do not have any dispute tag in their history. Moreover, there are known problems with these tags, such as issues of *disputes over tags* and *over-tagging*<sup>1</sup>. These issues arise when editors disagree on whether or not a tag should be added or removed from an article, and when different, possibly vague and non helpful tags are used for an article.

Finally, two previous studies [40, 76] considered a very limited set of tags (1 and 6 types of dispute tags, respectively), compared to the Wikipedia’s currently long and diverse list of *dispute templates*<sup>2</sup>: in at least in 16 of them, the words controversy and dispute are mentioned specifically, and others deal with less explicit forms of controversy. This shows that the tag taxonomies and their usage change over time, and also raises concerns to giving equal controversy weights to different tags whose intended meaning are hard to compare. For instance, it is hard to discern whether tags *Cite Check* and *Original Research* rise issues about controversial content, or trustworthiness of the content. Therefore, even though there might be different levels of controversy in different controversial articles, due to lack of reliable ground truth, we study the problem of identifying controversial articles regardless of their degree of controversy.

## 4.2.1 Classification vs. Ranking

By viewing controversy identification as a binary classification task, we need to convert the continuous scores obtained from some of the methods we study to a binary output. Scores in all of these methods are numeric, where higher values indicate more controversy. Mapping continuous outputs to binary outputs is a common task in many problems such as in diagnosing diseases based on one factor (i.e check [25], for instance.). More specifically, suppose  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  are series of examples, where  $x \in \chi$  is an instance represented by a continuous

---

<sup>1</sup>Wikipedia:Tagging\_pages\_for\_problems

<sup>2</sup>Wikipedia:Template\_messages/Disputes

score, and  $y \in \{-1, 1\}$  represents class labels (i.e. controversial or not). Now, assume  $f$  is a decision function that attaches a label  $y$  to each instance  $x$  as follows:

$$f_{\delta}(x) = \begin{cases} 1 & \text{if } x > \delta \\ -1 & \text{if } x \leq \delta \end{cases}$$

Then, the goal is to find the optimal threshold  $\delta$  that minimizes the misclassification error, which is equal to  $P(yf(x) \leq 0)$ . There are different classical methods for finding the optimal  $\delta$  such as grid search, ROC (Receiver Operating Characteristic) curves, and parametric models where a specific distribution such as normal distribution is assumed for the samples [25]. As the ROC curve is a more common method and does not make any assumption about the distribution of samples, we used this method in our work.

In the ROC curve approach, the optimal  $\delta$  is identified by varying the value of  $\delta$ , and calculating the true positive rate, and true negative rates for each value of  $\delta$ . Then, depending on the importance and weights of misclassifications of the two classes, the value of  $\delta$  that maximizes a combination of these two rates is chosen. For our problem, we assigned equal importance to the two classes (controversial and non-controversial), and thereby the optimal  $\delta$  is found for each score-based method at the threshold where the sum of true positive and true negative rates are maximized.

### 4.2.2 Metrics

We compare the methods using two criteria:

- Discriminative power: the accuracy of the method in separating controversial articles from non-controversial ones;
- Cost of training; which approximates the effort from the user before the method can be used.

### 4.2.3 Dataset

The dataset we used for our evaluation is the same as the dataset used in Chapter 3. Table 4.2 summarizes characteristics of this dataset.

Table 4.2: Statistics of datasets used for comparative study of controversy models

Category	Number	Strategy
Controversial	240	randomly chosen from all 15 different categories of topics
Non-controversial	240	100 random featured articles + 140 random other quality levels

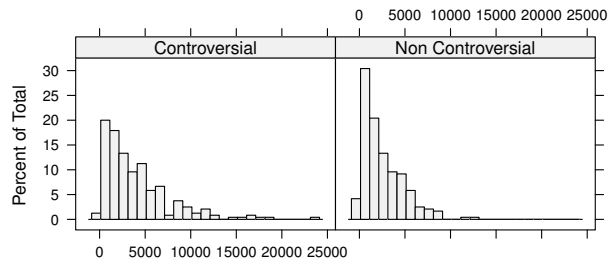
## 4.3 Experimental Results

### 4.3.1 Discrimination Power

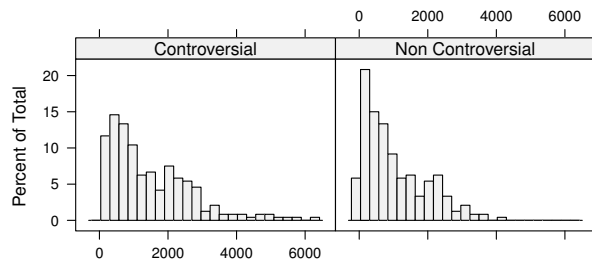
In this section, we compare different models in terms of their effectiveness in distinguishing controversial from non-controversial articles, which we refer to as the *discrimination power* of the methods. In addition, we considered some baseline methods to give better comparison of methods.

**Baselines** We considered the following intuitive baselines: (1) the number of unique editors contributing to each article (*#editors*); (2) the number of revisions of each article (*#revisions*); and (3) the number of revisions of discussion page associated with each article (*#talk-revisions*). Intuitively, one might expect that a large group of editors, a high number of revisions, or a long history of discussion should be indicative of controversy. As we show, even though controversial articles usually have large number of these factors, none of them alone is sufficient for telling controversial articles apart from others. For instance, long history of discussions is also common in featured articles, even though the goal is usually different from debating, and discussions are more centered around activities for improving the article coverage and style of writing.

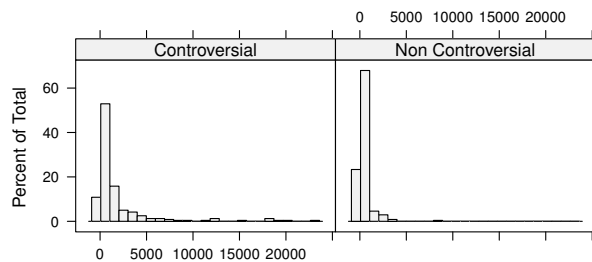
Figure 4.1 shows the distribution of controversial and non-controversial articles in our test set in terms of each of these factors. As can be seen, there are some examples of controversial articles with high values of the mentioned factors, but there are plenty of other examples that lay within the same range as non-controversial articles. Comparing the three factors, *#talk-revisions* is more discriminative than the other two, but still the samples of both classes are widely spread out and do not form a clear boundary. It should be noted that these three baselines are among the top-7



(a) Number of revisions



(b) Number of editors



(c) Number of revisions in talk page

Figure 4.1: Distribution of articles in our dataset in terms of baseline methods

ranked features in the Meta classifier we study in this chapter. In this method, Kittur et al. [40] combined these factors with other features in a classification approach to overcome the limited discrimination of each individual baseline.

**Metrics** We compared and evaluated the studied methods in terms of three metrics of: accuracy, precision and recall. Accuracy refers to the number of samples that were classified correctly. However, accuracy alone might not be sufficient in evaluating performance of a classifier, especially if we have highly unbalanced data across the two classes. Hence, we considered precision and recall as well, which are two common evaluation measures in information retrieval.

Table 4.3: Comparison of the studied methods in terms of accuracy, precision, and recall

Method	Accuracy	Precision	Recall
Mutual Reverts	0.67	0.60	0.55
Basic REA	0.60	0.56	0.83
Bipolarity	0.56	0.52	0.57
Meta classifier	0.75	0.73	0.86
Structure classifier	0.84	0.85	0.86
#talk-revisions	0.64	0.51	0.42
#article-revisions	0.57	0.53	0.46
#editors	0.56	0.56	0.45

In a classification task, precision for a modeled class is the number of correctly classified elements of the modeled class divided by the total number of elements that were classified as belonging to this class. Recall, on the other hand, corresponds to the number of correctly classified elements of the modeled class divided by the total number of elements that actually belong to this class.

In the context of our problem, the modeled class is controversial class, as our task is to identify controversial articles. In this context, precision means that how many of the samples that a classifier assigned them as being controversial were actually controversial. Similarly, recall means how many of the true controversial articles the classifier was able to find and assign them as being controversial. In this context, a system with high precision in correctly labeling controversial articles is desirable as it affects the experience and trust of its users (both readers and editors) who might rely on these labels when reading articles. On the other hand, recall can also be important, mostly for admins. For instance, admins might be interested in getting as many possible candidates of controversial articles as possible to be able to further investigate and assess these articles and manage to fix related possible issues.

**Results** Table 4.3 compares performance of the five studied methods along with the baselines based on accuracy, precision, and recall. The results are based on 10-fold cross validation experiment similar to experiment in Section 3.2.3.

The results show that the two classification-based methods have the best ac-

curacy. Moreover, there is a large gap between the performance of the Structure classifier and all the other methods. This highlights the importance of combining several different indicators and employing machine learning-based methods.

Also, note that #talk-versions is a baseline that has higher accuracy than some of the methods, such as Bipolarity and Basic REA. However, this is because this baseline classifies most of the instances as non-controversial, including a large fraction of controversial article which have short history for their discussion pages. Hence, we see that accuracy does not show the whole picture about the performance of methods. In particular, when considering the two other metrics, we see that all baselines have poor performance compared to most other methods, especially in terms of recall.

Also, some methods such as Basic REA and the meta classifier have very high recall score, but have much lower precision. In contrast, our Structure classifier is a very successful method that not only has the highest accuracy, but has both of its precision and recall among the highest values across all methods.

### **4.3.2 Cost of Training**

This section studies the effect of the amount of training data on the accuracy of the methods. The costs of collecting training data and training a model are usually very high as they typically involve human effort. Therefore, it is natural to seek trade-offs between accuracy and amount of training data.

It should be noted that the cost of applying a model is not limited only to the cost of providing training samples and can be extended to the cost of complexity and the availability of required resources to extract features and statistics related to that model. For instance, extracting a feature like the number of unique editors is much easier than features such as the number of articles linking to an article (as in the Meta classifier), or using a Wikipedia-specific resource such as election data to infer the attitudes (as in the Structure classifier). However, due to difficulty in objectively comparing these different factors, we can only analyze the cost from the perspective of number of training examples.

Unlike classification methods, the score-based methods do not require a train-



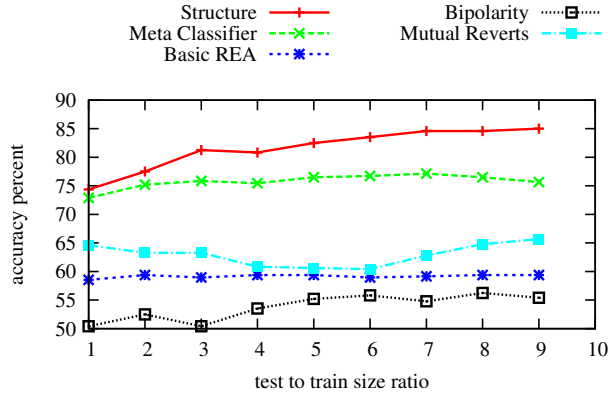


Figure 4.2: Effect of training size on accuracy

ing phase as they just assign a score to each sample. However, as explained in Section 4.2.1, in order to maximize the accuracy of these methods, an optimal cutoff value when scores are mapped to decision labels is needed. The optimal threshold found on a set of labeled data and used to label a set of unseen samples can differ from one sample data to another, which affects the accuracy of predicted labels on test data. Hence, we studied the effect of training sample size for these methods too.

**Results** Figure 4.2 shows the trend of accuracy of different methods when trained using data with increasing size, and tested on a fixed dataset. More specifically, we first partitioned the original dataset containing 480 articles into 90% for training and 10% as test data. Then, using that fixed 10% of test data, we tracked the accuracy of each model by training using only  $n * 10\%$  ( $\{n = 1, 2..10\}$ ) of the original training data. Finally, similar to the previous cross-validation experiment, to reduce variability of results, we did 10 rounds of partitioning the original data, where in each round, we chose a different partition as the test data, and considered the rest as the full training data. Therefore, the accuracy result of each training size was obtained by averaging results over the 10 rounds. Also, in generating a sample training at each training size and generating a test set at each round, we kept the same ratio of controversial and non controversial articles.

As can be seen, using more training data, overall, has a positive effect on the accuracy of all methods. However, relative benefits from using more training data dif-

fer across methods. For instance, the accuracy of the Structure classifier increases by more than 15%, while Basic REA has the least increase which is almost independent of the size of training data. In general as expected, the classification-based approaches are more sensitive to the amount of training data, while score-based methods show less than 5% difference for their results with the change of training size.

What is more interesting is that even when using just 10% of the available training data, both Structure and Meta classifiers achieved a very reasonable accuracy of about 75%, which is 10% higher than the best score-based method. Moreover, the relative performance of all methods remained the same, regardless of the amount of training data. This serves as a strong argument in favour of classification-based methods from the discrimination aspect.

## 4.4 Categorization of Controversy Models

In addition to score-based vs. classification-based categorization of controversy models discussed through out different experiments in previous sections, depending on the considered aspects and the resources used, these models can be further categorized into the following four groups:

- **Meta data-driven:** the methods in this group rely on extracting a set of numeric and simple statistics from the revision history of the article or/and its discussion page. These statistics are combined into a score or a set of feature vectors to be learned in a machine learning framework, or in a rule-based system.
- **User-driven:** in this category, controversy is modeled based on editors' interactions and their positive or negative collaborations. The Structure classifier and Bipolarity methods are examples of models in this category, which are based on signed network of collaboration of editors. Another example of methods in this category is the mutual reinforcing model of Vuong et al. [76], where the interaction of editors are modeled by the number of words they deleted from each other and the controversy of an article is calculated based

on an aggregation of the controversy scores of each two pairs of interacting editors (i.e. calculated in a recursive way).

- **Content-driven:** the third category of methods are those that model controversy by analyzing the content of revisions, comments, or the discussion pages. The content analysis can completely ignore the semantic of content of articles, and only apply simple content analysis such as tracking authorship and deleted words in the revision history of as in the Basic REA method. Alternatively, the content analysis can consider the semantic and apply Natural Language Processing techniques such as textual entailment of changed versions, or discourse analysis of the discussion pages. With some recent attempts on annotation of discussion pages [4, 64], these techniques seem to become more practical than before.
- **Pattern-driven:** the basis of methods in this group is analyzing patterns of edits over a history of revisions. The MR method that looks at mutual reverts in the revision history as sign of edit wars is an example of these methods. In a more advanced level, in a recent work, Wu et al. [81] modeled these edit patterns by network motifs, where the network motifs are defined by considering the network of editors and articles over each three consecutive revisions. In that work, they extracted the frequency of different network motif types (more than 39000 different types) over the entire revision history of articles as feature vectors and learned different edit patterns for controversial and non-controversial articles. Other edit patterns considering more abstract and general types, variable pattern length, and possibly unsupervised extraction of patterns can be studied in future.

Modeling controversy can also be improved by taking advantage of multiple categories and combining different sources. For instance, combining a meta classifier-based method with structure classifier as a user-driven method was shown to be superior to both of these individual methods as shown in Chapter 3. As another example, a user-driven model can be built by inferring the type of relations between editors based on a content-driven approach such as analyzing the comments

or discussions of the corresponding editors in discussion pages.

## 4.5 Conclusion

In this chapter, we studied five different controversy models in Wikipedia in terms of their discrimination power and the cost of learning the models. The results show that in practice the underlying principles of interaction of editors and the formation of controversy are too sophisticated to be captured by single heuristics. In particular, we showed that the three intuitive baselines of the number of revisions of the article, its number of editors, and the size of its discussion page, each alone, is not sufficient for detecting controversial articles. Hence, a combination of different factors need to be considered. In this regard, machine learning provides a suitable framework for learning the effect of different factors. In addition, as we showed machine learning methods have the advantage of being improved significantly with the usage of more training data, while maintaining a high performance even with small number of training examples compared to score-based methods.

On the other hand, score-based methods are easier to interpret, analyze and tune, especially when a more fine-grained analysis is needed. Examples of such analyses are ranking and comparing controversy across different articles or within different parts of the same article as explored in Chapter 6.

## Chapter 5

# Building PV Collaboration Networks

As discussed in Chapter 1, Wikipedia articles are built in a collaborative process where editors interact with each other all the time. These interactions can be positive indicating support and agreement, or negative indicating distrust and disagreement between editors. In the context of Wikipedia, a wide range of problems rely on some kind of modeling the type of interactions and collaboration relations of editors. Examples of these problems are assessing trustworthiness of articles [1,39,83], ranking editors based on quality of their past edits [12, 36, 42], analyzing controversy of articles [8, 40, 76], etc.

Determining the types of these collaboration relations in Wikipedia is challenging as they are not explicitly stated, unlike some other domains such as Epinions, Slashdot, etc. Instead in Wikipedia, the types of these relations have to be inferred from different edit actions of editors logged in the form of revision histories. For instance, Figure 5.1 shows a small fragment of the edit history of the article on Anarchism around March of 2006 (the “ $\Delta$ ” column indicates the net change in length of the article, measured in characters between consecutive revisions). Out of more than 15000 revisions that this article has, we focus on the interactions between two editors: *RJII*, who contributed 1,544 revisions, and *Infinity0*, who made 433 revisions. The disagreement between these editors is evidenced by their direct mutual accusations and the difference in their use of language: in this article, on average, *RJII* writes longer comments than *Infinity0* (70.6 characters vs 49.3 characters) and also uses more *positive* terms in his comments (423 versus 115). The sequence and timing of the actions is also revealing. The two editors are working concurrently,

time	editor	action	$\Delta$	comment
3:55	Infinity0	Rv	—	revert weasel words and pov
4:06	RJII	Rv	412	revert to rjii infinity is misleading the readers to think that tucker opposes employee employer relations...
4:09	Infinity0	Rv	-412	it says that tucker supported private mop please read your version uses many weasel words
4:12	RJII	Del	-131	anarcho capitalism tag
4:15	RJII	Ins	382	noting that tucker supports liberty of people to engage in employee employer relationships don't censor this fact
4:29	Infinity0	Del	-12	anarcho capitalism what's dubious it's a direct quote
5:21	Infinity0	Del	-264	anarcho capitalism
12:03	<i>other</i> <sup>†</sup>	Ins	41	ruined it

<sup>†</sup> Different user, with id VolatileChemical.

Figure 5.1: Partial edit history of article on Anarchism.

sometimes *fully* undoing each other's work (indicated as Rv or revert action in the Wikipedia logs) and and other times doing so *partially*, by deleting or inserting content to the previous versions (indicated as Del and Ins actions, respectively).

While the history snippet above is clear evidence that these two editors did not agree and collaborate with each other, it should be clear that analyzing the revision history of articles in search of sample agreements or disagreements would be virtually impossible for the reader. The sheer volume of data and the frequency with which the revision histories change make such an approach impractical. Moreover, not every editor writes descriptive comments. In fact, one can find several examples in further collaborations involving *Infinity0* in which he/she would simply revert back to a previous version without any justification. Besides, this example focuses on actions of two specific editors, while in practice, most of the time, we need to

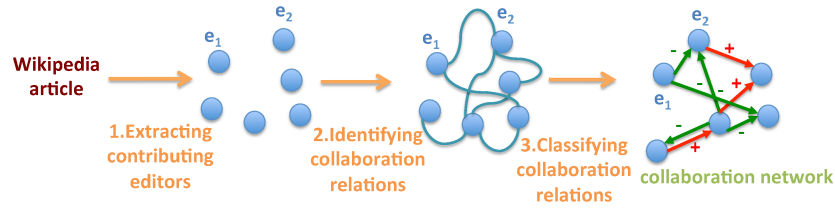


Figure 5.2: The workflow of building collaboration networks in our work

analyze an article from a more global perspective and by considering a larger set of editors.

Most previous works relied on simple statistics (often just counts) on these fundamental edit operations, such as “revert”/“delete” actions to measure disagreement, and “insert”/“restore” actions to measure agreement. While these were justifiable starting points, as revealed by our analysis, these simple statistics are not robust. Our aim in this chapter is to show how we can leverage history of collaborations of Wikipedia editors beyond these simple edit operations. Toward this goal, we show that by using a more global and extensive set of features that cover both individual and pairwise edit activities of editors, we can better model the true relationships among editors. In particular, we employ these relations in building PV networks which as shown in Chapters 3 and 4 resulted in significant improvement in identifying controversial articles compared to other methods.

In this chapter, we explain our method for inferring these relations and building PV networks based on them. To build these networks, we follow the workflow depicted in Figure 5.2. According to this workflow, we first build the set of nodes in the network of each article by extracting its contributing editors. Next, we identify which pair of editors have collaboration relation, and should be connected in the network of the given article. Finally, we assign positive or negative signs to edges connecting these editors by classifying their collaboration relation into corresponding classes. The details of these three steps are given in the following.

## 5.1 Extracting Contributing Editors

We define contribution of editors based on the number of revisions they edited in the given article. We found that about 50% of all editors in our corpus have only one edited revision across the entire history of the article. Hence, we consider contributing editors be those who have edited at least *two* edited revisions, and exclude all other editors from the set of nodes of network of each article to have a more manageable network size.

## 5.2 Identifying Collaboration Relations

There are different factors and properties that can be used to define collaboration relations between editors. We capture these relations based on the following definition:

**Definition 2** *Two editors are considered to have a collaboration relation if they have related revisions on the same article. The two editors are also considered to have an interaction for every such of those revisions.*

Perhaps the most intuitive way of determining that two revisions provided by different editors are related is to check whether they apply to the same text unit (e.g., the same sentence or paragraph) or, better yet, they concern the same issue within the article (e.g., both discuss the biography of the same person). While conceptually ideal, this notion of relatedness is not practical as one would need to manage, compare, and apply text understanding techniques to thousands of revisions, some of which modifying only a few words in the text. We follow a more pragmatic approach, which is based on the time lapse between revisions.

Intuitively, revisions that fall within a narrow window of time are more related as that implies that the later edit is triggered by a problem in the earlier revision. To account for different activity rates in different articles, we consider the number of revisions that fall between corresponding revisions of two edits instead of the actual time elapsed between them. Hence, for our purposes, the following gives a more clear and workable definition of relatedness.



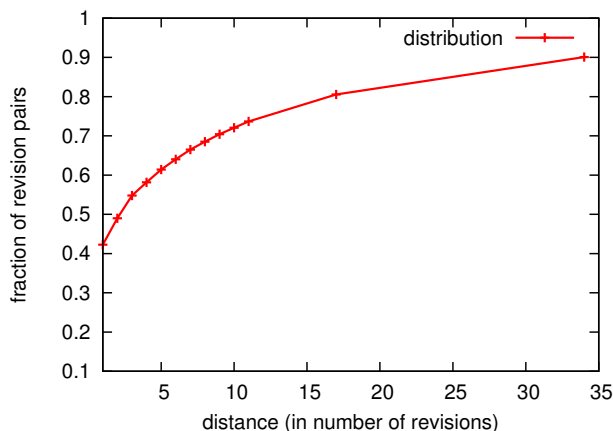


Figure 5.3: Distance distribution of revision pairs editing the same section

**Definition 3** *Two revisions of the same article are related if the number of revisions in between them is below a given threshold.*

In order to determine a reasonable threshold, we performed the following experiment on a random sample of 100 articles. We considered all pairs of revisions that modify the same *section*, and counted the number of revisions in between them. More specifically, for every revision  $r_i$  that edits the section  $s$  of article  $a$ , we record its revision distance to  $r_j$ , the first next revision after  $r_i$  that edits section  $s$ . We used sections for this experiment since each section is a moderate-size, independent conceptual unit (i.e. not too specific and small as a sentence, nor too broad and large as the whole article) that discusses the article from a particular aspect. Figure 5.3 shows the cumulative distribution of revisions according to distance. From the figure, we see that 40% of revisions affecting the same section are consecutive (and thus are clearly related), whereas over 70% of the revisions on the same section have at most nine revisions in between them. Our threshold was set at 34 revisions, which is sufficient to cover more than 90% of revisions on the same section.

### 5.3 Classifying Collaboration Relations

The final step in building our PV collaboration networks is to assign a positive or negative sign to the edges connecting editors who had a collaboration relation. For this assignment, we need to determine the type of collaboration relations of these

editors. To do so, we first collect information about each pair of connected editors in the form of a concise summary which is referred to as *collaboration profile*. We then classify the collaboration profile of each pair into positive or negative, and accordingly labelling the corresponding edge in the network. As training data, we need some reliable indication as to whether or not an editor trusts the opinion of other editors he/she collaborates with. As there is no specific training data for this task, we leverage the votes cast in the admin elections as a surrogate. These votes, along the collaboration profiles of editors are the primary sources in building PV collaboration networks, and therefore we called these types of networks PV networks as they are built based on profiles and votes. In the next two sections, we describe how we build these profiles, followed by our approach that uses votes for classifying them.

## 5.4 Building Collaboration Profiles

**Definition 4** A collaboration profile  $cp_{e_1, e_2}$  is a concise representation of individual and pairwise editing behavior of editors  $e_1$  and  $e_2$ , who have a collaboration relation.

For each collaboration profile,  $cp_{e_1, e_2}$ , a sign (positive or negative) will be assigned representing the sign of edges connecting editor  $e_1$  to editor  $e_2$ . We consider these edges to be directed as they denote attitude of editors towards one another, and attitudes are not necessarily symmetric. Hence, in the profile of  $cp_{e_1, e_2}$ , we refer to  $e_1$  and  $e_2$  as the *source* and the *target* editors respectively.

In building collaboration profiles, three categories of features are used as shown in Figure 5.4: a) individual features extracted from each editor’s edits, b) directional features derived from interactions and pairwise activities of two interacting editors in an ordered way showing behavior of one editor toward another, c) mutual features which similar to directional features are extracted from pairwise activities of editors, but are unordered. Note that each of the source and target editors have their own separate values for each of the individual and directional features, while they have a single, shared value for mutual features.

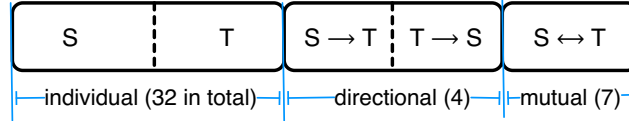


Figure 5.4: Three groups of features in collaboration profile representing the attitude of editor S (Source) towards T (Target)

Before we give more details about the features, we note that *edited articles* and *edited revisions* refer to only those articles and revisions contributed by the source or target editor in the collaboration profile. Next, we discuss the features in more detail.

### 5.4.1 Individual Features

Individual features provide a high-level description of the general type of articles edited, expertise and overall behavior of each editor. They are:

- number of articles edited (*articles*)
- number of revisions edited (*revisions*)
- average contribution size <sup>1</sup> over all edited articles (*contribution*)
- average concentration <sup>2</sup> ratio of all edited articles (*concentration*)
- number of agreement terms in comments of edited revisions (*agreement*)
- number of disagreement terms in comments of edited revisions (*disagreement*)
- number of agreement terms in comments of next revisions (*ag\_after*)
- number of disagreement terms in comments of next revisions (*dsg\_after*)
- number of times this editor reverted another revision (*revert*)
- number of times this editor restored another revision (*restore*)

<sup>1</sup>Contribution size of an editor in an article is the ratio of revisions made by the editor to all revisions made to the article.

<sup>2</sup>Concentration ratio of an article is defined as the ratio of unique editors to all revisions of that article.



from comments of 2000 revisions that are manually tagged as agreement or disagreement. Although not all examples of agreement or disagreement revisions may contain these terms, they can be a strong indication of agreement or disagreement between corresponding editors.

Example terms are shown in Figure 5.5 in the form of tag clouds. Specifically, looking at Figure 5.5a we see terms such as “fix”, “add”, “image”, “external”, “typo” which all are evident of edits with the intention of fixing or improving the article without drastically changing the meaning of its content. On the contrary, in the disagreement category, we see in Figure 5.5b that the most important terms suggest edits involving reverting a previous revision, fighting vandalism, referring to the talk page, removing the content, and finally clear opposition with the usage of words such as “uncited”, “irrelevant”, “do not”, “pov” (i.e. Point of View) and pronouns for addressing an editor(s).

In addition to comments, revert, restore, *Delta* size (where positive values are more likely to be inserts, and negative values to be deletes), and time difference between revisions are other indications of agreement and disagreement actions used in previous work [1, 8, 36, 42]. Note that all these behavioral features are individual features and indicate general positive or negative relation of the source (target) editor with respect to all other editors he/she had a collaboration with, and not only the target (source) editor in each collaboration profile.

Finally, the number of conflicting interactions happened in an article was considered as the conflict score of that article, which reflects the general type of the article with respect to controversy and dispute. In this work, we consider an interaction to be *conflicting* if the two corresponding revisions are:

1. consecutive and the later revision has an edit with negative *Delta* size, or
2. consecutive and the later revision has more disagreement terms than agreement terms, or
3. related and the later revision reverts the earlier revision.

## 5.4.2 Directional and Mutual Features

In contrast to the individual features, directional and mutual features focus on how the source and target editors in the profile behave with respect to each other. The directional features are:

- ratio of co-edited articles to edited articles (*coed\_d\_arts*)
- ratio of edited revisions in co-edited articles to all edited revisions (*coed\_d\_r\_evs*)

The mutual features are:

- number of co-edited articles (*coed\_arts*)
- number of interactions (*interactions*)
- fraction of conflicting interactions (*conflicting\_interactions*)
- fraction of interactions in consecutive revisions (*consecutive\_interactions*)
- average revision distance between all interactions (*interaction\_distance*)
- average concentration score of co-edited articles (*coed\_concentration*)
- average conflict score of co-edited articles (*coed\_conflict*)

The intuition behind relying on these features is that the higher the number of co-edited articles and interactions between the two editors, the more information is available about how they treated each others revisions, and the easier their attitude can be identified. Also, the shorter interaction distance is, the more we can be certain that the source editor reacted in response to the target editor's edit. Moreover, this distance might provide a distinction between negative interactions and positive interactions. Similarly, the percentage of interactions between the editors that were either conflicting or consecutive (thus, with a higher chance of being responded) helps capture the overall attitude of one editor towards another.

## 5.5 Classifying Collaboration Profiles

Given the profiles of editors and their collaborations, our goal is to classify each collaboration into one of *agreement* or *disagreement*. In the absence of labelled data, one needs to resort to heuristics to infer labels for collaborations. For instance, Maniu et al. [53] used features such as the number of deleted, inserted and replaced words and whether an editor has given barn-star award to another editor, and label each feature intuitively as a sign of a positive or negative relation. Then, the final sign of the relation of a pair of editors is determined based on the sign of the majority class. Bogdanov et al. [5] developed a content-based method by building a topic model of edits. In their method, the relation of a pair of editors editing the same paragraph of an article takes a value in the range  $[-1,1]$  depending on whether one editor changes the topic distribution of the paragraph towards the changes made earlier by the other editor of that paragraph or not. This approach again relies on a heuristic which is limited to interactions that can change the topic distribution of an article; the method also has not been completely evaluated and is only shown to be useful for two articles as case-studies.

### 5.5.1 Leveraging Admin Elections

We take a more systematic approach by leveraging the strong relation that exists between the way Wikipedia editors collaborate in editing articles, and the way they later vote in admin elections.

The intuition behind using admin elections as our training data is that an editor who casts, for example, a negative vote to a candidate is more likely to have a negative than a positive interaction with that candidate before casting his vote. Admin elections have been used by Maniu et al. [53] with votes being a deciding feature in the sign of relation between two editors. However, they could only use this feature for pairs of editors who participate in elections, and the number of such pairs is much smaller than the number of interacting editors.

We use the election data to learn the weight of features that contribute to positive or negative collaborations. More specifically, we use the election data and tag a

limited set of interactions as positive or negative; a classifier is built on this labelled data, which can then be used to predict the sign of collaboration profiles for other editors who may or may not appear in the election dataset. Our results show that such a classifier can achieve high accuracy, which supports our observation that past iterations are highly influential on the attitudes that editors have towards one another, and thus on how they cast their votes.

To build and train this classifier, for each candidate  $c$  and voter  $v$  who appeared in an election, we first build their collaboration profile  $cp_{v,c}$ . This profile is used then as a feature vector for one training sample, and the sign of the vote is considered as its corresponding training label. Using all the collected votes and feature vectors, we train a classifier called “vote classifier”. We use this classifier to infer the attitudes of all interacting editors and assign signs of edges in PV networks.

## 5.6 Experimental Results

In Chapter 3, we used the task of identifying controversial articles as a successful application of PV networks. In that chapter, we argued that the success of PV networks greatly depends on the method used for assigning the sign of edges connecting editors who had collaboration relation. In this section, we evaluate the accuracy of the vote classifier, which in turn determines how accurate our PV networks are.

### 5.6.1 Dataset Description

The election data is available in Wikipedia dump in the form of special articles, named “Request for Adminship” (RFA). We collected and parsed all these RFA articles from a Wikipedia dump (date April 5, 2011), resulting in a dataset that covered 3713 elections. More statistics about this dataset is shown in Table 5.1.

We use election data to train vote classifier using collaboration profiles as feature vectors. Hence, we can only predict those sets of votes, where we could build a collaboration profile for the candidate and the voter. This requirement caused to be able to assess vote classifier on a smaller dataset compared to original election dataset we extracted. We refer to this smaller dataset as “mapped election” data,



Table 5.1: Statistics of election dataset

number of elections	3713
number of unique editors	9541
positive votes	130193
negative votes	36239

Table 5.2: Statistics of the extracted, and the mapped election data

	extracted data	mapped data
total	166432	89652
positive	130193	75168
negative	36239	14484

compared to original “extracted election” data. Table 5.2 shows number of votes across extracted and mapped datasets, along with break-down to positive and negative votes. As we can see from this table, the ratio of positive votes is 78% and 83% in extracted and mapped datasets respectively.

It should be noted that in predicting the sign of votes, we followed these two guidelines: First, to simulate the real time situations and to predict the sign of the votes before they are cast, we use only the information that is available prior to each vote. Second, a candidate and a voter can appear in multiple elections possibly at different times and the vote of the candidate can change from one election to next. In cases where  $v$  casts a vote for  $c$  only once in the entire election dataset, all revisions up to the time of casting vote seem to be relevant and are used for building collaboration profiles. Similarly, in cases where  $v$  casts the same vote for  $c$  multiple times, all revision history up to the time of vote is considered. However, for multiple conflicting votes, we consider only the revision history from the time of the most recent previous conflicting vote  $v$  casts for  $c$ .

## 5.6.2 Overall Prediction performance

Table 5.3 shows the performance results of vote classifier on predicting votes using a 10-fold cross validation experiment on on two datasets: *full* and *balanced*. The

Table 5.3: Results of predicting votes from collaboration profiles on full and balanced dataset in terms of accuracy (Acc) and area under ROC curve (AUC)

Model	Full-Acc.	Full-AUC	Balanced-Acc.	Balanced-AUC
Random Forest	0.869	0.877	0.781	0.857
J48	0.842	0.706	0.695	0.707
SMO	0.838	0.5	0.579	0.579
Logistic	0.837	0.626	0.591	0.628
All positive	0.838	0.5	0.5	0.5
Avg-positive	0.339	0.496	0.5	0.5

full dataset contains all training data we could obtain. The balanced dataset is obtained from the full dataset by randomly sub-sampling positive votes until the number of positive and negative votes are the same. For these results, we tested four classifiers, namely Random Forest, J48, SMO and Logistic using their default settings in Weka <sup>4</sup> machine learning tool.

In our experiments, we found that the Random Forest classifier achieves the highest accuracy among the studied classifiers in both datasets. In fact, Random Forest classifiers have a good performance in general and also on imbalanced datasets as shown in some previous work [38] due to their bagging and internal feature selection methods. Hence, for this classifier, we applied an additional tuning and feature selection method by following the approach proposed by Reif et al. [63]. In particular, for ranking and selecting features, we used the Gini importance metric of the classifier, and removed 5 features with the lowest importance score. These features were 1) number of co-edited articles, 2) fraction of conflicting interactions, 3) fraction of interactions in consecutive revisions, 4) voter’s average contribution size, and 5) ratio of co-edited articles to all articles edited by voter. After selecting features, we tuned the two parameters of the classifier which led us to choose 70 trees and 15 random features at each branch for training the classifier.

This additional tuning and feature selection resulted into 86.9% and 78.1% accuracy, an about 1% improvement (which translates to over 1000 more correct prediction in our dataset) and 8% over the default setting of Random Forest in Weka on

<sup>4</sup>[www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)

Table 5.4: Top-15 important features of vote classifier in predicting votes

# of candidate’s agreement terms ( $agreement_{candidate}$ )
# of candidate’s revisions ( $revisions_{candidate}$ )
candidate’s avg. contrib. size ( $contrib_{candidate}$ )
candidate’s avg. time being responded ( $time\_after_{candidate}$ )
# of candidate’s disagreement terms ( $disagreement_{candidate}$ )
avg. $\Delta$ size of edits after candidate’s ( $Delta\_after_{candidate}$ )
avg. $\Delta$ size of candidate’s ( $Delta_{candidate}$ )
candidate’s avg. response time ( $time_{candidate}$ )
# of candidate’s reverted revisions( $revert\_to_{candidate}$ )
avg $\Delta$ size of voter ( $Delta_{voter}$ )
# of candidate’s edited articles ( $articles_{candidate}$ )
# of reverts made by the candidate ( $revert_{candidate}$ )
avg concentration ratio in articles edited by candidate ( $concentration_{candidate}$ )
avg concentration ratio in articles edited by voter ( $concentration_{voter}$ )
avg concentration ratio in co-edited articles ( $coed\_concentration$ )

full and balanced datasets respectively. Table 5.4 shows the top 15 features ranked by importance metric of Random Forest classifier.

As is shown, the features that are ranked on top are mostly individual features, and are that of the candidate; this is consistent with our intuition that the activities of the candidate are more influential on the outcome of a vote than the characteristics of the voter. The top 15 features includes from interaction features only the average contribution size of co-edited articles which is representative of how collaboration work is divided between editors of each article, and whether most of the edits are done by a few editors, or a large number of editors are involved in revising the articles.

### 5.6.3 Comparison with other methods

We also compared our vote classifier with two simple baselines: all-positive and avg-positive; the former classifies all votes as positive whereas the latter classifies a vote as positive if the number of edited revisions of the corresponding candidate is higher than the average (number of edited revisions of all candidates), and negative otherwise.

Using the number of edited revisions of candidates as a threshold in the avg-positive baseline is based on the intuition that the previous number of edits of a candidate is an important factor and a low number is sometimes cited as the reason for negative votes in our dataset.

Comparing all these methods on the full dataset, we can see that our best results using the Random Forest classifier shows about 3% and 37% improvement over the strong all-positive baseline in terms of respectively accuracy and area under ROC curve (which is a measure commonly used for imbalanced datasets [38]). On balanced dataset, the difference of our method and baselines is even more visible, where our accuracy is 28% higher than the best baseline.

## 5.7 Conclusion

In this chapter, we proposed a novel method for inferring attitudes of Wikipedia editors towards one another, and building PV collaboration networks as an efficient type of collaboration networks. Our method is based on correlating previous collaboration history of editors to how they vote for each other in admin elections. Based on this assumption, we were able to train a classifier for not only inferring the type of collaborations of candidate and voter editors, but for any pair of editors who had a collaboration relation while editing articles.

We showed the effectiveness of our attitude inference method as a main component of PV networks in Chapter 3 for identifying controversial articles. In this chapter, we further evaluated our inference method by verifying performance of the vote classifier in predicting the signs of votes. As features, we used only previous history of collaborations of editors in the form of collaboration profiles, which consisted of an extensive set of individual and pairwise collaboration features. The results of our experiments on a dataset of more than 89000 votes on a very unbalanced dataset, and more than 14000 votes on a balanced dataset showed efficiency of this classifier in correctly predicting the sign of votes. This confirmed our intuition of influence of collaboration history of editors on some of other types of social relations of editors in Wikipedia.

## Chapter 6

# Fine-grained Analysis of Controversy at Text-unit level

Automatically determining whether an article is, or has been, controversial helps Wikipedia readers as well as admins, by warning them about the potential bias and imbalance in the coverage of the article. Another useful task would be to determine the specific parts such as sections or paragraphs within each article that are responsible for most of the disputes and conflicts between editors. This is because in a manual inspection of some controversial articles, we observed that controversy can be often attributed to specific parts of the article, which are the focus of most of debates and disputes between editors. For instance, sections about *abortion and breast cancer* and *questioning the authorship of some of the works attributed to Shakespeare* were one of main reasons of conflicts in the *Abortion* and *Shakespeare* articles respectively. These observations are also consistent with the hypothesis raised by Li et al. [45]. Identifying and highlighting these specific parts, referred to as *text-units*, help readers to have better understanding about these controversial articles and to be able to separate the disputed parts of the articles from other reliable and accepted parts.

Hence, in this chapter, to give more insight about controversial articles, we analyze controversy at a finer level than the whole-article level described in previous chapters. We refer to this problem as *unit-level analysis*, and we approach it using an optimization objective that considers the effect of the text-units on overall controversy of the article. We show this optimization problem can be solved efficiently

if this effect is modeled in a way that satisfies the two conditions of *monotonicity* and *submodularity*. Hence, we explore different methods for defining such models, and discuss possible ways for evaluating them. We show that not only designing these models is challenging, but there are several difficulties for evaluation of this problem that overall prevented us from being able to fully solve this problem.

## 6.1 Problem Formulation

We formulate the unit-level analysis based on an unsupervised selection of the top text-units having the most *contribution* in making the article controversial. More specifically, let  $U$  be the set of all text-units appeared at some point in the history of the article  $a$ , and  $F : 2^U \rightarrow \mathbb{R}^{\geq 0}$  be a function modeling contribution of a set of text-units of  $a$ . Then, we are interested in finding a set  $S \subset U$  containing  $k$  text-units that overall has the most contribution in controversy of the article  $a$ .  $k$  is a parameter given by the user depending on how many text-units are desired to be seen in the output.

In the simplest case, text-units can be assumed to be independent and thereby contribution of a set of text-units can be modeled by the sum of the contribution of each unit:

$$F(S) = \sum_{u \in S} F(u)$$

. In this case, the solution can be obtained by examining each unit separately, ranking them in terms of their contribution values, and selecting the top- $k$  units (i.e. we can also stop before  $k$  steps if at step  $i < k$  we reach a unit with zero contribution as inclusion of units after  $i$  step does not change  $F(S)$  anymore.). However, without this independence assumption, we will have the more general version of the problem, where the objective is to maximize  $F(S)$  as follows.

$$S^* = \operatorname{argmax}_{S \subset U} F(S) \quad \text{subject to : } |S| \leq k \quad (6.1)$$

where  $S^*$  is the optimal set of  $k$  units,  $U$  is the set of all units that appeared at some point in the history of the article, and  $F$  is a function measuring contribution

of units. We also assume that an empty set does not have any contribution, and thereby  $F(\emptyset) = 0$ .

### 6.1.1 Computational Complexity

The optimization in Equation 6.1 is an NP-hard problem in general [57, 80]. However, as discussed by Nemhauser et al. [57], when  $F$  is *submodular* and *monotone*, a greedy algorithm can find an approximate solution whose  $F$  is guaranteed to be at least  $\frac{e-1}{e} \sim 0.63$  of the optimal solution. The greedy algorithm in each step selects a unit  $u$  which provides the highest positive difference between the value of the function before and after selecting this unit:  $F(S \cup \{u\}) - F(S)$ . If this difference becomes zero, then the algorithm stops before  $k$  steps and the output contains less than  $k$  units.

Monotonicity holds if  $\forall A \subset B, F(A) \leq F(B)$ . On the other hand, submodular functions are those that satisfy a property referred to as “diminishing returns”, where the effect of adding a new item like  $u$  to a set of already selected items like  $S$  decreases as the set  $S$  grows. More formally, one of the definitions for submodularity is to have  $F(A \cup \{u\}) - F(A) \geq F(B \cup \{u\}) - F(B)$ , where  $A \subset B \subset U \setminus u$  ( $U$  is the set of all items). When this equation is satisfied everywhere with the equal case, the function  $F$  is called modular, and  $F(A) = c + \sum F(u)$ . This case is in fact the same as assuming independence of the items and having a linear contribution score as explained before. Hence, the independence case is a special case, which can be addressed by the greedy method. Moreover, the greedy algorithm required to solve this case can be based on only evaluation of  $F(u), \forall u \in U$ , and hence the optimization will be the same as ranking of units. The proof is provided in the Appendix.

Note that both submodularity and monotonicity are consistent with the intuition one might have about controversy. For instance, we expect that a set of units to have higher or at least equal contribution value compared to a smaller subset of them. This is because each additional unit, in the worst case, can have zero contribution to the controversy of its owner article, but never can decrease the global controversy of it or have negative contribution. Similarly for submodularity, it is

expected that adding a new unit to a set of text-units will change contribution of the set more when the set contains fewer units. The intuition in this case is that when more text-units have been selected before, it is more likely that they already have captured all controversial points of an article and leave no room for contribution of new text-units.

## 6.2 Defining Contribution Function

We consider two approaches for defining the function  $F$ , where they measure the effect of inclusion or exclusion of a set of text-units on the global degree of controversy of the article. More specifically, let  $f : A \rightarrow \mathbb{R}^{\geq 0}$  be a function assigning zero or positive controversy score to each Wikipedia article. Then, in the first approach we model contribution of a set of text-units  $S$  of article  $a$  by considering

$$F(S) = f(S)^1, \tag{6.2}$$

which states that these text-units contributed to controversy of article  $a$  by as much as much controversy score of this article would be if it had contained only text-units  $S$ .

On the other hand, in the second approach we model this contribution by considering

$$F(S) = |f(a) - f(a - S)|^2 \tag{6.3}$$

where  $f(a - S)$  stands for controversy of article  $a$  after removal of units  $S$ . In this way, we assume that the set  $S$  contributed by as much as the removal of this set changes controversy score of the article compared to when all of its text-units are included.

We refer to these two approaches as *inclusion* and *exclusion* models respectively.

---

<sup>1</sup>Note that the set of text-units can be considered to be like a synthetic article containing only those revisions changing units in that set, and hence  $f$  can be applied to it.

<sup>2</sup>The operator absolute value is used to avoid having negative contribution value. Hence, this model measures the absolute relative change in controversy of an article after removal of a set of units.



### 6.2.1 Computing with Revision History

The complexity of calculating  $F(S)$  according to the described inclusion or exclusion models depends on the choice of controversy function  $f$ . Let us represent article  $a$  with its list of revisions  $R = \{r_1, r_2, \dots, r_n\}$ . Each  $r_i \in R$  is also represented by a tuple of  $r_i = \{ed_i, ts_i, cm_i, tx_i\}$ , where  $ed_i$ ,  $ts_i$ ,  $cm_i$  and  $tx_i$  are the contributing editor, timestamp, comment and the text of the article after applying this revision respectively. Then applying  $F(S)$  requires us to obtain  $a'$ , the modified version of article  $a$ , which is done by only considering text-units  $S$  out of all text-units of  $a$ . Considering only this set of text-units makes some of the revisions of  $a$  to be removed completely from  $a'$ , while making some to appear with some changes.

For instance, assume a simple controversy function where  $f$  measures the number of revisions of the article:  $f(a) = \#revisions(a)$ . With the inclusion model, we will have  $F(S) = \#revisions(a')$ . In this way, if revision  $r_i$  does not change any of units  $u \in S$ , its text will become the same as the text of its previous revision and thereby it becomes a duplicate revision. Hence, we can exclude it from revisions of  $a'$  as it does not apply any change to  $a'$ . On the other hand, if  $r_i$  involves some changes to unit  $u \in S$ ,  $r_i$  can be seen to be modified to  $r'_i$ , where all of its components are the same as  $r_i$ , except its  $tx$  component. The text of this modified revision,  $tx'_i$ , will be obtained from  $tx_i$  by considering only units  $S$ . Figure 6.1 shows two examples of obtaining  $r'_i$  from  $r_i$ . Similarly, when exclusion model is considered, first a set of units  $S' = U - S$  is obtained and then  $f$  is applied, considering only  $S'$  units similar to inclusion model. Finally, the value of  $|f(a) - f(S')|$  is calculated as the final value of  $F(S)$ .

Most of the controversy models discussed in Chapter 4 also work by only considering some information extracted from the revision history of the article. Hence, obtaining  $a'$  for them is similar and follows the same principle. We refer to these controversy models as content-only models.

In contrast, when the controversy model uses some information beyond the article's history of revision, this task will become more complex. For instance, the Meta classifier described in Chapter 4 extracts some meta information from the dis-

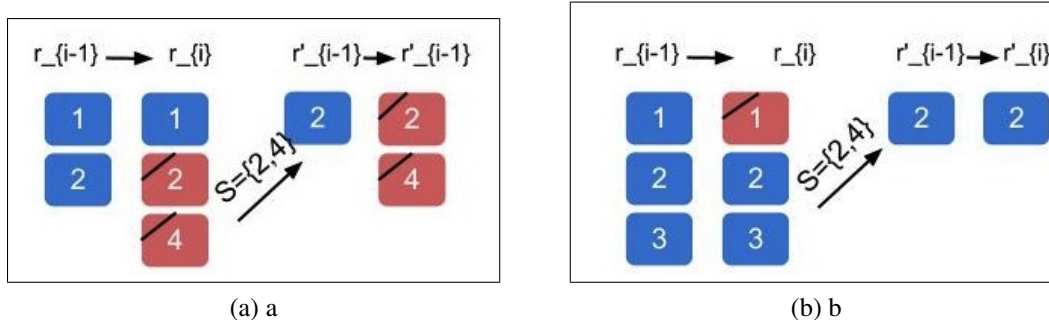


Figure 6.1: Obtaining modified revisions under the inclusion model, where in (a)  $r_i$  changes some units from  $S$ , while in (b) it does not change any of units of  $S$  and thereby  $r'_i$  becomes a duplicate revision. Units in “red” (having a diagonal line) are those that have been changed from revision  $i - 1$  to revision  $i$ , and those in “blue” are those that have not changed between these consecutive revisions.

discussion page of the article as well. As inclusion or exclusion of a set of text-units can affect the topics discussed on the discussion page, the features related to this part should also be updated in applying  $F(S)$ . However, updating these features is challenging as there is no clear mapping between the topics discussed in the discussion page and the text-units found on the article itself.

Therefore, we see that the two considered contribution models are suitable when used with a content-only controversy  $f$ , which includes most of the previous controversy functions studied in Chapter 4.

## 6.2.2 Assessing Current Controversy Models

In this part, we aim to test existing controversy models experimentally against the desired computational properties discussed before to see whether any of them can be used as contribution function  $F$ . Although, in practice, analytical evaluations are required to be able to demonstrate suitability of a method with respect to these properties, the experimental results can still give some insights about each method. We focus on monotonicity property in this part, as the analysis pertaining to submodularity would be similar.

In order to evaluate different controversy models, we used 50 controversial articles randomly chosen from the set of 240 controversial articles, using sections as text-units. For each article, we rank the units according to the number of edits they

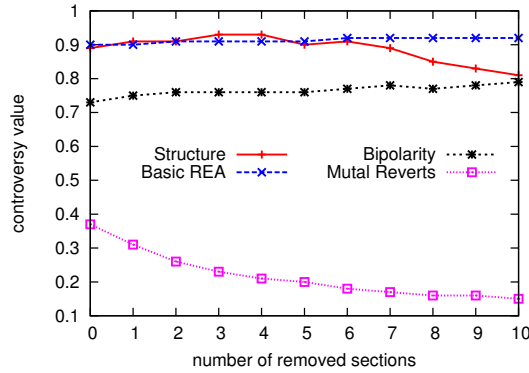


Figure 6.2: Monotonicity test for the studied methods

comprise. Then, we successively remove units in decreasing ranking, and measure the controversy score obtained from each method after each removal, recalculating all features as necessary. In this way, at each step, the controversy function corresponding to each method is applied on a set of text-units that are subset of text-units of the previous step. Hence, to have a monotone contribution function based on a controversy model using one of the two approaches of inclusion or exclusion, that controversy model should produce scores that decrease with removal of units.

We excluded the Meta classifier from the experiments due to difficulties in calculating contribution function  $F$  as explained before. We also turned the Structure classifier into a controversy measure by using the membership probability of the controversial class as controversy score (easily derived from Weka).

### 6.2.2.1 Discussion

Figure 6.2 shows the changes in absolute controversy scores of different methods as more and more text-units are removed. As we can see, Mutual Reverts is the only monotonic method. With the removal of more text-units, the total number of disagreement edges in networks of Bipolarity method decreases, but the way this change affects the structure of the networks, in most cases cause the bipolarity score to fluctuate in both directions. Our Structure classifier does not seem strictly monotonic either.

Based on these experiments, we see that designing controversy scores in a way to be useful for defining contribution of units and solving the problem of selecting

the most contributing text-units is challenging. On one hand, we have classification-based methods such as Structure and Meta classifiers that model controversy well at the article level, but normally it is difficult to have the desired computational properties with them. Note that, in general, it is possible to build a classifier having the desired properties. As an example, Lin et al. [46] discussed building a monotone submodular classifier based on using simpler submodular and monotone functions (i.e. for instance the number of nodes in our Structure classifier, or the number of revisions in Meta classifier are submodular, and monotone functions) as features and learning to combine them using only positive weights.

However, the main problem with classification-based methods is sparsity in the training set. More specifically, to convert the classifier decision to a controversy score, one should rely on class membership probability values in assigning instances to the controversial class. These probability values, when applied at the level of units to calculate  $f(a - S)$  or  $f(S)$ , might be unreliable. This is because these instances, specially when working with small number of text-units (obtained by either inclusion or exclusion of some units), can get very far away from the space of the original instances used to train  $f$  at the whole- article level. In this case, the assumption of modeling class membership and controversy score based on probabilities obtained from the classifier will not be correct.

On the other hand, we have score-based methods such as Mutual Reverts that satisfy the required computational properties, but they assign scores which are not consistent with Wikipedia controversy labels at the whole-article level, and therefore cannot be good candidates for defining contribution of units within articles. Hence, future work can focus on designing methods having both acceptable article-level controversy model and computational properties, or coming up with completely different ways to define contribution of text-units.

### **6.3 Evaluation**

Evaluation methods for finding controversy at the text-unit level is challenging as it is a novel problem and to the best of our knowledge, except Baykau et al. [10],

no work has been done on it before. However, as discussed by Dori-Hacohen [18], Baykau et. al [10] evaluated their system using list of Latest Edit Wars of Wikipedia, which are not the same as controversial articles we study. Also, we are not aware of any labelled data that could be used for a complete evaluation of this problem.

The first option for evaluation of unit-level analysis is to use a small sample of articles tagged specifically with a section-level dispute tag. There are two controversy related dispute tags at section level: `{{POV section}}` and `{{disputed section}}`, where each has 766 and 368 articles having at least one section with these tags respectively. From this large number of articles, we found that only 13 of them are among the list of manually tagged controversial articles. This small number of articles can be used to evaluate methods for the case of  $k = 1$ , where we are interested in selecting the single top unit that contributed most to controversy of the article. However, beyond this case and for a more thorough evaluation, we need a human judgment experiment.

### **6.3.1 Difficulties of a Human Judgment Experiment**

For a human judgment experiment, one may aim at building a ground truth by asking annotators to annotate a small number of articles by choosing a set of sections in each article that they think have the most contribution in controversy of that article. However, even assuming that sections are independent text-units with separate contributions to make the problem easier, the task is too complex and expensive as a human judgment experiment. In particular, each annotator would have to do the whole task individually by rating the controversy degree of each section and ranking all sections according to his/her ratings and it is not possible obtain such ratings by breaking the task to simpler sub-tasks done by different annotators. Coming up with these ratings would require that an annotator check the entire revision history of the article containing all edits applied to different sections, along possibly considering discussions of the discussion page. With average of 50 sections, and 5000 revisions per article the task would be impractical.

Alternatively, one may consider a simpler annotation task by asking annotators to compare 2-3 sets of text-units (sections or paragraphs), where one set is chosen

by the proposed method and the others are chosen by some baseline methods. In this way, annotators would have to test only text-units in the given sets instead of testing all text-units of each article. However, even in this setting, the annotation experiment would be very time-consuming and expensive as annotators still need to go through all revisions corresponding to each text-unit in each given set. As text-units with the highest controversy contribution are likely to be among the most edited text-units in each article, it is expected that testing each single text-unit will take about an hour. Therefore, for a very basic setting such as two annotators, 50 sample articles, two test methods, and sets of size three for each method, we need about 600 annotation hours. Assuming we need to pay only 10\$ for each hour of work, this annotation task will become very expensive and time-consuming. In addition, this estimated cost and time is for when the annotation experiment is done only once. However, in practice multiple rounds of test and experiment might be needed to train annotators for the given task and learn to how design the experiment to increase annotators agreement. Also, the proposed method might not work efficiently with the first attempt and might need some improvements and fixes that applying them forces to have another round of annotation for the improved method.

## 6.4 Conclusion

Analyzing controversy in Wikipedia can be improved by determining sources of controversy in each controversial article. Towards this goal, in this chapter, we introduced the unit-level analysis problem, where the goal is to locate the text-units having the most contribution in making an article controversial. We formalized this problem using a combinatorial optimization problem, and discussed its complexity and some circumstances under which it has tractable approximation schemes.

A natural way to define contribution of a set of text-units of an article is to study the their effect (either by including or excluding them) on the global controversy of the article. Hence, we experimentally examined current controversy models to determine whether they satisfy monotonicity, which is one of the necessary condi-

tions for tractability. We showed that this is the case only for a simple score-based method, which does not perform very well at the controversy detection task. Future work is needed therefore, with focus on designing new controversy models that are amenable for this finer-grained analysis or finding other ways to define contribution of text-units. Alternatively, further algorithmic development, leading to other bounded approximation schemes for the problem might be needed.

Finally, aside from the choice of the contribution function, we also discussed the challenges of evaluating the unit-level problem, which necessitates building a benchmark of known issues related to a set of controversial topics, or finding other ways for providing a finer-level analysis of controversy. In the next chapter, we describe another approach towards this goal.

## Chapter 7

# Fine-grained Analysis of Controversy at Revision level

In Chapter 6, we discussed the text-unit analysis as one approach for providing a fine-grained analysis of controversy in Wikipedia. In this chapter, we describe another approach, which is referred to as *revision-level* analysis. This analysis works at the level of revisions of each article and aims to identify the revisions that contributed most to controversy of that article.

Finding these revisions helps readers and editors to grasp a better knowledge about controversial points and opposing views. This is because disputed issues and opposing views may have been expressed in older versions of the article and might now be hidden from the attention of a reader who usually looks only at the current or the most recent versions of the article. Manually searching for these issues in the revision history is virtually impossible as controversial articles usually have a long history of revisions, where revisions containing disagreements are mixed with peaceful (i.e. just applying normal, collaborative changes such as improve text style or adding a new information) and vandalism revisions. Hence, highlighting these specific disagreement-involved revisions helps to obtain a summary of the most important disputed issues that resulted in conflicts between editors. For instance, linking abortion to breast cancer is one of the most disputed issues in “Abortion” article. Finding revisions debating this issue helps to identify this topic as one of the disputed issues in this article. Moreover, it gives some insights about how this article evolved in regard to this issue, and how much of the arguments and opinions



of each opposing side has been reflected at each particular state of this evolution. A similar situation can be seen for “Osama bin Laden” article, where there is a long history of disputes on linking this person to “September 11 attack”.

For this purpose, in this chapter, we first formulate this analysis using the same optimization objective introduced in Chapter 6. We then describe our solution, which is a submodular and monotone function developed based on *maximum coverage problem*. We show effectiveness of this proposed method compared to other methods using two different evaluation measures. Finally, we show its usefulness qualitatively by some case studies.

## 7.1 Problem Formulation

The revision-level analysis and the unit-level analysis introduced in Chapter 6 are similar problems, where the goal in both of them is to select a set of article elements that have the most contribution in making an article controversial. These article elements in the case of unit-level analysis correspond to text-units, while they refer to revisions in the revision-level analysis. Hence, for this analysis, we can use the same maximization objective shown in Equation 6.1.

For convenience, we repeat this equation here:

$$S^* = \operatorname{argmax}_{S \subset U} F(S) \quad \text{subject to : } |S| \leq k \quad (7.1)$$

This time, the reference set  $U$  will become all the revisions in the history of the article, which are represented by  $R = \{r_1, r_2, \dots, r_n\}$ . Each revision  $r_i \in R$  is also shown by a tuple of  $(tx_i, ed_i, et_i, cm_i, ts_i)$ , where they stand for text, edit actions, editor, comment and the timestamp of the revision respectively. Also,  $tx_i$  can be seen as a list <sup>1</sup> of terms  $T_i = (t_1, t_2, t_3, \dots, t_n)$ , where terms are the smallest text-units such as words, phrases, or topics for representing text of the article. Furthermore,  $ed_i$  is the list of edit actions represented by  $ed_i = ed_i^+ + ed_i^-$ , where  $ed_i^+$  and  $ed_i^-$  are lists of inserted and deleted terms in  $r_i$  respectively. These lists

---

<sup>1</sup>Note that we used *list* and not *set* as a term might appear multiple times in different parts of the article and get edited more than once

are obtained by comparing the list of terms in  $tx_i$  and  $tx_{i+1}$  using a Diff algorithm. Finally, it is assumed that  $r_0 = (\emptyset, \emptyset, null, null, null)$ .

Based on these set of revisions, we need to define the contribution function  $F$  in a way to reflect the effect each revision will make on future revisions of the article. Intuitively, we expect higher contribution for a revision that applies an edit to the text of the article that causes further conflicts and disputes between editors, compared to, for instance, a revision added by a vandal editor that gets undone very soon. Also, a revision that introduces a modification that gets accepted by other editors and does not change further is expected to have little or no contribution at all .

Similar to the problem of selecting text-units, we might first think of defining contribution function based on article-level controversy models. However, a controversy contribution model defined that way will not produce reliable contribution values for very small sets as usually controversy models are not able to quantify controversy of the article using only a few revisions. In particular, current controversy models are not applicable to a single revision. Therefore, we will have difficulties with these models in having meaningful values for  $F(e), e \in U$ . Hence, we followed another approach based on the standard framework of maximum coverage problem as explained next.

## 7.2 Coverage-based Contribution Function

### 7.2.1 Background

The maximum coverage problem is a well-known problem in approximation algorithm theory, which has been used in many summarization works before [46,69,70,73]. The goal in these works is to summarize a document (or a corpus) by selecting a set of  $k$  sentences (or documents) from it in a way to maximize the coverage of information contained in that document (or corpus). In these works, information is usually approximated using basic text-units such as single words, phrases or topics of the documents. We refer to these text-units as *terms* in this chapter.

Maximizing coverage of these terms drives to select terms that are not already

covered, and therefore provides a natural way to promote diversity of the selected sentences (documents), which is a necessary property for a summary. Also, while different objectives have been used to model coverage of information in different works, one common way for representing them is as follows:

$$C(S) = \sum_t \theta(t) g_{d \in S}(\phi(d, t)) \quad (7.2)$$

where  $t$  is a term from the set of terms used for representing ideas expressed in documents  $S$ . Also,  $\theta$  is the importance of  $t$  in the entire set of documents, and  $\phi$  is the importance of this term in a particular document  $d$  from the set of documents  $S$ . These two importance metrics are referred to as *global* and *local* scores respectively. Finally,  $g$  is a function that integrates individual local scores of documents  $d \in S$  for all terms  $t$  that are common across the set  $S$ . Usually this function is chosen in a way to make the entire coverage function  $C$  to be submodular. The rationale behind choosing a submodular  $C$  function is to have the sub-optimality guarantee of maximizing submodular functions discussed before [57].

### 7.2.2 Adapting to Our Problem

We employed the general objective model of Equation 7.2 in our contribution function by changing the definition of global and local scores in a way to not maximize the coverage of important ideas as in summarization works, but rather to maximize the coverage of *controversial* ideas. In other words, we select the most contributing revisions by selecting those that maximize the coverage of controversial ideas, where similar to summarization works, ideas are approximated with terms. In particular, we adapt the general coverage function to our problem as follows:

$$F(S) = \sum_{t \in e_S} \theta(t) \sum_{r_i \in S} \phi(r_i, t) \quad (7.3)$$

where  $e_S$  is the list of edited terms in the set  $S$ , which is obtained by getting the union of all  $e_i$  corresponding to  $r_i \in S$ . Hence, for each  $r_i \in S$ , we only focus on terms that appear in its edit actions part (i.e.  $e_i$ ). Also, we considered terms to be only common noun phrases and named entities (the three groups of persons,

Table 7.1: List of top-20 terms in “Abortion” article

abortion	pregnancy	induced abortion	method	issue
abortion law	fetus	trimester	united states	pregnant woman
woman	case	number	risk	uterus
pro-life	miscarriage	procedure	death	link

organizations, and places), as they determine the general topic of the controversy. They also have been used in many opinion mining works as the opinion holders and targets [13, 50, 58].

As for function  $g$  in Equation 7.2, we used the  $\sum$  function assuming independent contribution for different revisions (i.e.  $F(r_i \cup r_j) = F(r_i) + F(r_j)$ ). This choice makes our overall function  $F$  to be modular, where the objective function introduced in Equation 6.1 can be solved optimally. In the next two subsections, we describe how we implement  $\theta$  and  $\phi$  functions, which correspond to term-global and term-local scores respectively.

### 7.2.3 Defining Term-global Score

We define the global score of terms according to Equation 7.4.

$$\theta(t) = change(t) + mention(t) \quad (7.4)$$

In this equation,  $change(t)$  is the relative number of times that term  $t$  is inserted or deleted in the entire revision history of the article (i.e. relative with respect to the total number of revisions of the article). Similarly,  $mention(t)$  is the relative number of times that this term is mentioned in the discussion page of the article (i.e. similarly, relative with respect to the total number of revisions in the discussion page of the article). The combination of these two factors gives an estimate of the controversy score of this term at the global level. The intuition here is that a term that is used more in the discussion page and appears more in the edits applied to text of the article is more likely to be part of disputes and conflict between editors.

Table 7.1 shows the top-20 terms for “Abortion” article based on global-term scores. As can be seen, most of these terms are general topic phrases that can be found on disputes about this topic.

## 7.2.4 Defining Term-Local Score

We define the term-local score based on Equation 7.5.

$$\phi(r_i, t) = \begin{cases} 0 & |e_i| > \delta \\ \sum_{sec \in Sec_{r_i}} \phi(sec, t) & else \end{cases} \quad (7.5)$$

In this equation,  $|e_i|$  is the edit size of revision  $r_i$  (i.e. measured in terms of the number of terms in  $e_i$ ). If this size is greater than the threshold  $\delta$ ,  $\phi(r_i, t)$  is considered to be 0 leading to have no contribution for revision  $r_i$  regardless of term  $t$ . This step is considered to prevent having large contribution scores for those vandalism-related revisions that involve big edits such as mass-deletes.

For other cases,  $\phi(r_i, t)$  is calculated based on the sum of the local scores of term  $t$  in  $Sec_{r_i}$ , which is the set of all sections that were edited in revision  $r_i$ . We map scores of terms in each revision to their scores in sections they got edited, as each term can appear multiple times in a revision and under different contexts. For this purpose, we consider section  $sec$  in revision  $r_i$  to be a tuple of  $(tx_{sec}, ed_{sec}, et_{sec}, cm_{sec}, ts_{sec})$ , where all components have the same value as in  $r_i$  if this section has been edited in  $r_i$ , and will be null otherwise. This local score at the section-level, in turn, is calculated as in Equation 7.6.

$$\phi(sec, t) = \frac{1}{|N_i|} \sum_{sec' \in N_i} sim(sec, sec') + ds_g(sec') + rsp(sec, sec') \quad (7.6)$$

In this equation,  $N_i$  is the set of  $n$  revisions after revision  $i$ , and  $sec'$  is the match of section  $sec$  in each of these revisions. This match only exists if revision  $r_j \in N_i$  contains an edited section whose title or text is sufficiently similar to title or text of section  $sec$ . The parameter  $n$  specifies a window, where we study the impact of the changes applied to section  $sec$  in revision in  $r_i$  on its future revisions. Smaller values assign high scores to changes that cause short-lived sparks of disagreements, while larger values assign high scores to only long-lasting disagreements. We set this window size according to editors collaboration window size of Chapter 5.

Also, the three functions of  $sim$ ,  $ds_g$  and  $rsp$  model the likelihood of seeing conflicts and disagreements between editors while editing section  $sec$ . More specifically, the first function,  $sim(sec, sec')$ , measures similarity of the terms changed in

each of the sections  $sec$  and  $sec'$  respectively. The intuition is that often in controversial articles a series of edits in the form of edit-war is observed, where the same or very similar text is removed and then reintroduced multiple times in a sequence of revisions by different opposing editors. Hence, high similarity of the edited text applied on sections  $sec$  and  $sec'$  can be indicative of such edit-war patterns. We used the ratio of the number of common terms of  $sec$  and  $sec'$  to the number of terms of  $sec$  as the similarity of these two sections.

The second function,  $dsg(sec')$  is the ratio of disagreement words appearing in the  $cm_{sec'}$ , the comment part of section  $sec'$ . For extracting disagreement words, we rely on our manually compiled list of words explained in Chapter 5. Seeing disagreement words in a comment of a revision can show the negative attitude and disagreement of the editor towards previous edits, and hence  $dsg(sec')$  shows how much disagreement we will see on editing section  $sec$  after  $r_i$ '.

Finally, the third function that we consider for calculating local score at the section-level is the function  $rsp(sec, sec')$ , which returns 1 if the last authors editing sections  $sec$  and  $sec'$  are the same, and 0 otherwise. In this way, we measure how many times the author of section  $sec$  *responds* to other edits on this section by *coming back* to edit it again in subsequent revisions.

Combining all these three functions gives us a local score for each section  $sec$  edited in revision  $r_i$ . This score will be used as the term-local score of all terms edited in this section, as they are considered to belong to the same context and issue. Then, as explained, these section-level local scores will be combined to get revision-level local scores for each edited term.

### 7.2.5 Summary of the proposed Revision Selection Method

Figure 7.1 shows a simple example of calculating  $F$  for revision  $r_i$ , which contains two edited terms:  $t_1$ , and  $t_2$ . According to Equation 7.3,  $F(r_i)$  depends on global and local scores of the edited terms. For global scores, we need to calculate  $\theta(t_1)$  and  $\theta(t_2)$  using Equation 7.4. These two functions depend on the total number of times that each of these terms has been edited in the revisions history of the article, or has been mentioned in its discussion page.

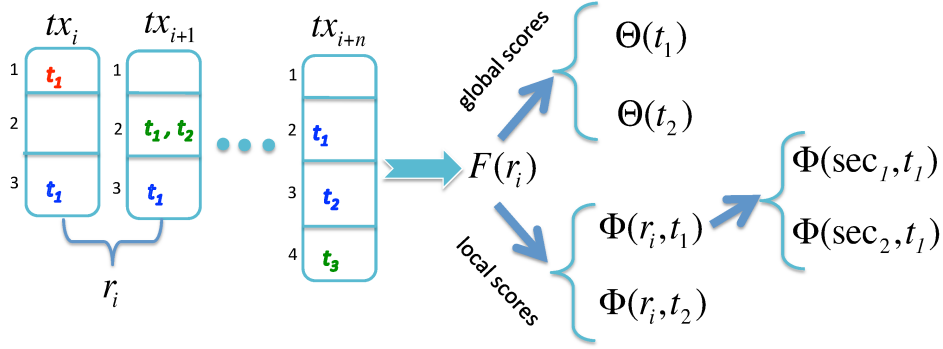


Figure 7.1: An example of calculating term-global and term-local scores. Inserted, deleted and unchanged terms are shown in green, red and blue respectively.

On the other hand, for local scores, we need to calculate  $\phi(r_i, t_1)$  and  $\phi(r_i, t_2)$  based on Equation 7.5. According to this equation, the local score of a term in a revision depends on its local scores across all sections it has been edited. For instance, as shown in Figure 7.1, term  $t_1$  has been edited in sections  $sec_1$  and  $sec_2$ . Hence, we need to calculate  $\phi(sec_1, t_1)$  and  $\phi(sec_2, t_1)$  using Equation 7.6, which tracks different evidences of conflicts across the next  $n$  revisions after  $r_i$ . In the next step, the value of these two functions will be summed up to get the revision-level score of term  $t_1$ :  $\phi(r_i, t_1) = \phi(sec_1, t_1) + \phi(sec_2, t_1)$  as this term has been edited in both  $sec_1$  and  $sec_2$ .

Combining local and global scores of all edited terms in  $r_i$  gives us the final  $F$  value of this revision, which is used to assess its individual contribution value.

After calculating the individual contribution value of each revision in an article, we can select the set of  $k$  revisions that are believed to have the most contribution in making the article controversial. As  $F$  is defined based on the coverage function  $C$  of Equation 7.2, it satisfies submodularity and monotonicity properties. Hence, the sub-optimality guarantee that exists with these properties can be leveraged to choose the top- $k$  revisions in a greedy way. In fact, our function is a specific type of coverage functions that is modular as we used a modular function for its  $\phi$  component. This allows us to be able to solve the optimization equation introduced in Equation 6.1 optimally and much faster, compared to sub-optimal solutions that can be obtained for general submodular functions. For this purpose, we only need to

rank revisions of each article by their calculated  $F$  values in decreasing order and choose the top- $k$  revisions.

We refer to our proposed coverage-based method for selecting revisions as Coverage-based Controversy Contribution (CCC) model and compare it with some baseline methods as explained in Section 7.5.

## 7.3 Evaluation

Evaluation of the problem of selecting revisions similar to the problem of selecting text-units is challenging as there is no labelled data for it. For this problem, the first option we considered was to use dispute tags. These tags are added to a specific revision of an article once dispute was observed by editors and usually are left in subsequent revisions until dispute gets resolved. Hence, one might think that revisions containing dispute tags should have more contributions in controversy of the article compared to other revisions in the other parts of the history of article lacking these tags. However, upon further investigation, we found out that we cannot rely on these tags for evaluation as they are usually added sometime after the dispute has already started, and might not be removed immediately after the dispute has been resolved. Besides, there might be several normal and vandalism revisions applied to the article in the same time period that the article has dispute tags.

Hence, we adopted two other approaches which evaluate the selected revisions at two different levels of “individual”, and “set” levels, as explained next.

### 7.3.1 Set-level Evaluation

At this level, we measure the quality of the selected revisions as a whole by measuring the global controversy of each article one time using all of its revisions and one time with all revisions excluding these selected revisions. We then show that after each exclusion, on average controversy of articles drop more when the excluded revisions are selected by our method, suggesting that these revisions actually contributed more to controversy.

For measuring controversy of articles, we need to have a discriminative and



monotone controversy measure that not only be accurate according to Wikipedia controversy labels, but can also show the changes in controversy of articles after excluding selected revisions in a monotone and consistent way. For this purpose, we chose *Mutual Reverts* method as this method is the only method satisfying monotonicity according to results shown in Section 6.2.2 (analytical proof is also provided in the Appendix). This method is not as accurate as some of the methods we studied in Chapter 4. However, as shown in that chapter, it still has a reasonable accuracy and it is a score that has correlation with controversy degree of articles.

More specifically, we use Equation 7.7 to measure the drop of controversy of articles after removing the selected revisions. We refer to this quantity as “MR-drop”.

$$MR - drop(a) = \frac{MR(a) - MR(a')}{MR(a)} \quad (7.7)$$

where,  $MR(a)$  and  $MR(a')$  are controversy scores of the article  $a$  (i.e. calculated using Mutual Reverts) before and after removing the selected revisions respectively.

### 7.3.2 Individual-level Evaluation

We also evaluate the selected revisions individually to see how much each selected revision is likely to have contributed to controversy of its article. For this purpose, we relied on the edit-classifier developed by Daxenberger et al. [15], where a taxonomy of edits containing 21 different classes was defined. Based on this taxonomy, Daxenberger et al. developed a multi-label classifier to assign one or more classes to each edit (i.e. each revision is mapped to a set of edits which are obtained by the Diff algorithm) in each revision of a Wikipedia article. The authors trained this edit-classifier on a manually-developed corpus of 1995 edits, and reported to achieve 0.67 in terms of overall F-measure and above 0.70 for most classes.

#### 7.3.2.1 Mapping edit-classifier labels to controversy labels

Daxenberger et al. also mapped these 21 classes to 3 high-level categories in their work:

- Text-base edits: these edits contain changes that change the meaning of the content of the article, by inserting, changing or deleting the *text* (i.e. information), *references*, *files* or the *templates* of the article.
- Surface edits: these edits contain changes that do not change the meaning of the content of article, and include edits such as spelling and grammar corrections, paraphrases, relocations, etc.
- Policy edits: these edits are edits that apply a change that is categorized as vandalism or reverting an earlier edit.

Based on this high-level categorization, we expect that category of surface edits to correspond to edits that do not involve changing the article in a way to cause more controversy. In contrast, category of text-base edits is more likely to contain such edits, even though not all classes in this category can be considered to imply disagreement and conflicts between editors. For instance, the class “information-insert” in this category can correspond to cases where an editor inserts new information to an article in a way that gets accepted by other editors later. Also, classes related to edits of files, and templates in this category are more likely to be minor issues affecting the style and presentation of the article, rather than changing its main information. Therefore, from category of text-base edits, we only considered the two classes of “information-delete” and “information-change” as controversy contributing classes.

Also, the edit-classifier was reported to have poor performance for the class of “vandalism”. Hence, we ignored this class in this category from our analysis.

We then applied the edit-classifier on each of the revisions selected by each of the studied methods. As explained before, the edit-classifier works at the level of edits within each revision, while the methods we evaluate work at the whole revision-level. Hence, we mapped edit labels obtained from the edit-classifier to revision labels by assigning the corresponding class label to each revision if it contains at least one edit with such label.

### 7.3.2.2 Evaluation metrics

We then evaluated different methods using two metrics from the information retrieval works: Precision at  $k$ , and Mean Average Precision (MAP). Precision at  $k$  is simply the fraction of relevant documents in the top  $k$  retrieved results, but does not consider the position of the relevant documents within the obtained results. Hence, it only shows the overall quality of the results.

MAP, on the other hand, is the mean of precision computed at every relevant document, and considers order of relevant documents by favouring the results where relevant documents are shown first. It is the most commonly used single-value summary of the performance of ranking systems [54].

In the context of our problem, the task of retrieving documents corresponds to selecting revisions according to each studied method. In this context, we considered a revision to be relevant if it has been assigned to at least one of the two classes of “information-change”, or “information-delete”, and irrelevant otherwise. While the set of controversy contributing revisions might not be limited to these two classes, we expect revisions from these classes to be more likely to have contributed to controversy than revisions from other classes that merely involve minor edits, fixing spelling errors, or adding new information. Hence, a method that selects more revisions from these classes is expected to have better performance than a method that selects more revisions from other classes.

## 7.4 Comparison with Other Methods

We compare our method with four different methods:

1. *Random*: In this method, we select revisions randomly.
2. *Edit-size*: In this method, we select revisions by the size of their edits and first choose those revisions containing the biggest edits. The intuition for this baseline is that the bigger the size of edit of a revision is, the more deviation from the previous revisions has been introduced, and consequently the more contribution to controversy might be observed.

3. *Diversity*: In this method, we select revisions in a way to promote diversity of the selected revisions as the more diverse the selected revisions are, the more they are likely to contain the contrasting and opposing edits contributed to controversy of the article. For diversity we used the diversity metric defined in [75]. This metric is defined based on the entropy of the set of items (i.e. revisions in our case) as follows:

$$H(S) = - \sum_{i=0}^n p(f_i) \log(p(f_i)) \quad (7.8)$$

where  $f$  is a random variable that has  $n$  different values, and  $p(f_i)$  is the probability that the value of this feature to be  $f_i$  given the instances in set  $S$ . Higher values of  $H(S)$  correspond to more diversity in the features present in instances of set  $S$ .

In our implementation, we considered three types of features: a content feature which is extracted from the text of revisions; an author feature which corresponds to the set of all authors contributing to the article; and a time period feature which is obtained by partitioning the revision history of the article to equal parts of size  $n$ . We arbitrarily set the value of  $n$  to be 100, having  $h/100 + 1$  different time period features, where  $h$  is the size of the revision history of the article.

## 7.5 Experimental Results

In this section, we report the results of comparison of our method against the other three baselines described in Section 7.4. Except the Random baseline, in all methods, we need to represent revisions by the set of terms edited in them. We obtained these terms by extracting named entities and noun phrases of the text ( $tx$ ) or edited part ( $ed$ ) of each revision using Stanford Named Entity Recognizer<sup>2</sup> and the LBJ Part of Speech Tagger<sup>3</sup> respectively. Also, for the Random baseline results are obtained by averaging over 10 runs with different seeds.

<sup>2</sup><http://nlp.stanford.edu/software/CRF-NER.shtml>

<sup>3</sup>[http://cogcomp.cs.illinois.edu/page/software\\_view/POS](http://cogcomp.cs.illinois.edu/page/software_view/POS)

After this step, we calculated scores of all revisions in each article according to each method, and then proceeded with selecting the best  $k$  revisions for each method. This selection for all methods, except the Diversity method requires only to rank revisions and choose the top- $k$  revisions. However, the Diversity method is a submodular method and the set of best  $k$  revisions can only be approximated using a greedy algorithm as explained in Chapter 6. For this purpose, we used the CELF lazy algorithm proposed in [44]. The main idea in this algorithm is to choose the element  $e$  (i.e. revision in our problem) at each step that maximizes  $\delta(e) = F(S \cup e) - F(S)$ , while using the lazy evaluation idea to recalculate the value of  $\delta(e)$  only when it is necessary. This is based on the fact that when the function is monotone,  $\delta(e)$  does not need to be calculated for all elements, and many of the elements can be naturally filtered out. This leads to fewer function evaluation and causes dramatic speedup compared to the original greedy algorithm [44], while not affecting the returned selected elements.

### 7.5.1 Set-level Results

Figure 7.2 shows the results of comparing different methods in terms of the drop in MR score. As can be seen, with selecting more revisions, the MR score keeps dropping for all methods. This is consistent with the monotonic nature of the MR score explained in Section 6.2.2. However, what is important is that this drop is significantly higher for CCC method compared to other methods across almost all numbers of selected revisions. In particular, the difference of CCC and other baselines is more visible when fewer revisions have been selected. For instance, when only 10 revisions are selected, we see an MR-drop of about 0.10 for our method. This means that on average the MR score of each article in our test set decreases by about 10% of its original MR score. On the other hand, Random and Diversity methods have drop of less than 0.03 at this level. Also, the Edit-size baseline, while ties with our method at some values, overall shows less MR-drop for both low and high ranges of the number of selected revisions.

This suggests that CCC method is able to select revisions that are more likely to have contributed to the controversy of controversial articles as removing them

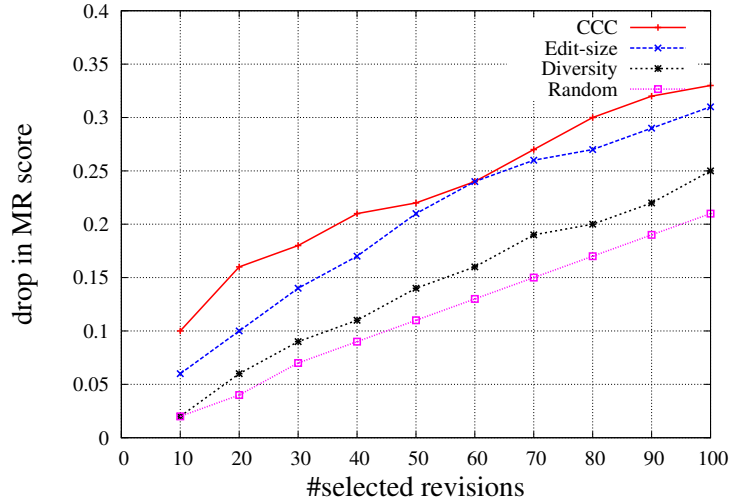


Figure 7.2: MR-drop of different selection methods at different numbers of selected revisions

causes to see more drop in the overall controversy of these articles. Specially, CCC shows more difference compared to other methods when only a few revisions are selected from each article. These revisions have more effect on the experience of users, as users usually are only interested in investigating the top selected revisions.

## 7.5.2 Individual-level Results

Figure 7.3 shows the results of comparison of methods in terms of precision at  $k$ . As can be seen, the studied methods have the same performance order similar to the MR drop experiment, where CCC method achieves the best results, and Edit-size is the strongest baseline among the competitive methods.

Also, similar to that experiment, the difference of CCC method and other methods is more when less revisions are selected. For instance, for 10 revisions, the precision of CCC is about 9% higher than the best baseline method.

Another point to note in this experiment is that that overall the precision of all methods is low. This is because we have a very restrictive definition for relevant revisions based on considering only revisions with class labels of “information-delete” or “information-change”. While there might be other possible classes for a contributing revision, as explained in Section 7.3, we considered only these two specific classes because they are more reliable to distinguish controversy-related

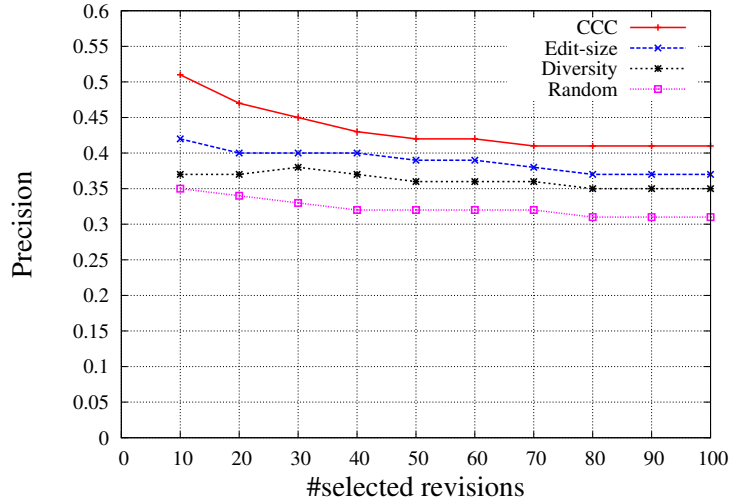


Figure 7.3: Precision at  $k$  of different selection methods measured by using edit-classifier of Daxenberger et al. [15]

CCC	Edit-size	Diversity	Random
<b>0.27</b>	0.24	0.17	0.15

Table 7.2: Results of comparison of methods in terms of MAP

revisions from normal or vandalism-related ones.

Next, we compared different methods in terms of MAP. The results of this comparison are shown in Table 7.2, where we see that CCC achieved the highest value. Also, we applied t-test and found that the difference of this method and each other method is statistically significant at  $p=0.01$  level using two-tailed t-test.

In summary, these results show that CCC not only is able to select more relevant revisions at different sizes of selected set, but it also selects these relevant revisions at higher up in its list of selected revisions. Therefore, it is more useful in helping users to investigate these relevant revisions.

### 7.5.3 Detailed Examples

In this section, we examine the revisions selected by the CCC method for “Abortion” and “Osama bin Laden” articles. We selected these articles as our examples because they are about familiar and well-known controversial topics. They are also highly-debated in Wikipedia as evidenced by long history of discussion in their dis-

cussion pages, and the fact that some specific articles have been created to address the conflicting points separately (i.e. for instance, the articles “Abortion debates” and “CIAal-Qaeda controversy” are about some of the controversial issues on each of these articles respectively).

Tables 7.3 and 7.4 show the top-10 selected revisions for “Abortion” and “Osama bin Laden” articles respectively. In these tables, index and id refer to the position of the selected revision in the revision history of the article, and its unique revision identifier assigned by Wikipedia’s system respectively <sup>4</sup>. We also provided a brief description about the intention and related topic of each selected revisions, where we manually extracted this information from the edit actions and comments. Next, we explain what can be found about these selected revisions.

### **7.5.3.1 Revisions with comments**

As can be seen, the comments of several of selected revisions contain clear disagreement and strong opposing views with respect to previous revisions. For instance, the author of revision of the 2nd selected revisions in the Abortion article has commented “reverting to medically accurate facts that are not sanitized to be PC”. In this revision, this author changed the article back to include the phrase “associated with the death of the human [[embryo]] or [[fetus]].” in the introduction section of the article which contains definition of the abortion. Whether this phrase should be included in the definition of the abortion or not was the focus of many debates and discussion in the discussion page of this article. Also, the author of this particular revision pushed his point of view several times during the revision history of the article, and his edits were accused to be POV (i.e. edits violating the Neutral Point of View policy of Wikipedia) by several other editors.

As another example, we see that the 1st selected revision in “Osama bin Laden” article contains a comment that clearly speaks about one of the conflicting issues in this article. This issue which caused multiple opposing edits and discussions is about whether or not Osama bin Laden was involved in the September 11 attack.

---

<sup>4</sup>Information about each revision such as its editor, edit actions, etc. can be found from [https://en.wikipedia.org/w/index.php?diff=prev&oldid=\[revision-id\]](https://en.wikipedia.org/w/index.php?diff=prev&oldid=[revision-id])



### **7.5.3.2 Revisions without comments**

Among the selected revisions, we also see revisions without any comments. However, tracking edits and changes in the subsequent revisions confirms that these revisions also contributed to controversy of their articles. For instance, the 1st and 4th selected revisions in Abortion article have no comments, but looking at the applied edits and subsequent revisions shows us that they are about the conflicting issues of definition of abortion which explained before. This issue is highly debated in this article as evidenced by multiple related edits and discussion in the discussion page of the article. For instance, only searching the word “fetus” or “embryo” in the discussion page returns more than 50 results, where editors have long debates and discussions on related topics such as whether fetus can be considered as a human-being or not, whether it is a medical term that should be used in the first section of this article or not, whether abortion should be just defined as termination of pregnancy or it should be mentioned as well that it causes the death of the fetus, etc. Similar evidences can be also found for the 2nd, 4th and 5th revisions selected for “Osama bin Laden” article that link this person to September 11 attack.

### **7.5.3.3 Related revisions**

Another point to notice is that a set of related revisions can be seen in the selected revisions of both articles. These related revisions are about the same disputed issues that were raised during a specific time period of the history of the article. For instance, the 1st, 2nd, 4th and 5th selected revisions in “Abortion” article are nearby revisions that happened within a window of 5 revisions and are all about the issue of definition of abortion. In this article, the 6th and 7th are also both about the issue of father’s right.

A somewhat similar situation can be seen in “Osama bin Laden” article, where the 4th-7th revisions are nearby and related to linking bin Laden to September 11 attack.

The selection of these related revisions can be attributed to the fact that we used a modular contribution function  $F$ , where contribution of different revisions are assumed to be independent. However, in some applications, these related revisions

might be seen as redundant, and it might be more preferable to get a set of more diverse results that cover different disputed issues. In this way, a submodular contribution function that assigns higher scores to set of revisions that cover different controversial terms should be employed instead.

#### **7.5.3.4 Irrelevant revisions**

Finally, we see that the 8th and 9th revisions of the “Osama bin Laden” article are clearly not relevant to creating controversy in this article. These revisions are marked with a “\*” in Table 7.4. The 8th selected revision is a short-lived edit-war between two editors who disagree on a minor issue related to whether the last name of “Osama bin Laden” should be used to refer to him in subsequent sections of the article, or whether he should be simply referred to as “Osama”. This shows that our heuristic and, in particular, the window of tracking the impact of a revision are not suitable for this example, and we need more tuning to be able to handle these cases correctly. On the other hand, the 9th selected revision is a clear vandalism revision, where its editor changed many of the words of the article with vulgar and nonsense words such as changing “Osama bin Laden” to “Osama **pig** Laden”. This shows that despite some of the heuristics we employed for assigning low scores to vandalism revisions, our method still can fail for some cases. This necessitates using more sophisticated heuristics for such revisions or completely first filtering them out by means of some of the recent developed vandalism classifiers [37, 55, 61, 79]. However, it should be noted that in general vandalism detection is a hard problem and current methods are far from perfect [61].

#### **7.5.3.5 Summary**

Overall, these detailed examples show that our method is able to highlight some of the most important disputed issues in the studied articles, even when only a few revisions are selected. Such selected revisions help users to grasp a high-level knowledge about controversial articles and can work as a starting point for seeking more information about them. For instance, a web-based user-interface can be built to show the selected revisions as a summary of the most important changes in a con-

Index	Id	Comment	Description
1235	14346854	-	changing the vocabulary used for defining abortion
1238	14347417	<i>reverting to medically accurate facts that are not sanitized to be PC</i>	changing the vocabulary used for defining abortion
5429	47657778	<i>restoring last consensus version prior to protection</i>	changing definition of abortion and also changing some parts about ABC hypothesis
1240	14347644	-	changing the vocabulary used for defining Abortion
1237	14347389	<i>I'm reverting this because I feel the edits to be POV and misleading (the morning-after pill cannot cause an abortion, so don't use the verb abort</i>	changing the vocabulary used for defining Abortion, and also changing some parts about morning-after pills
723	12059682	-	arguments about father's right
720	12050881	-	same as the previous selected revision
5431	47658296	<i>restoring - only four people have agreed that there is consensus on the new opening - see main talk page</i>	changing definition of abortion and also changing some parts about ABC hypothesis
5541	486263980	<i>less bias and more background in start</i>	changing the vocabulary used for defining abortion
3379	32955737	<i>clarify abortion risks</i>	add some information about risks of abortion which are disputed in a few of later revisions

Table 7.3: Top-10 selected revisions for Abortion article

controversial article. In its simplest form, such a user-interface can show the selected revisions as a list of revision-ids that each linked to the corresponding version of the article once clicked. In this way, the user can see the text of the article in a corresponding contributing revision, along being able to see its related information such as the set of applied edits, timestamp, editor, comment, etc.

## 7.6 Conclusion

In this chapter, we introduced another approach for analyzing controversy in Wikipedia, which worked at the revision-level. In this approach, our goal was to highlight a set

Index	Id	Comment	Description
7874	85998551	<i>no, that is pure conjecture, we have no evidence and even FBI spokesman this last summer said we have no hard evidence</i>	linking bin Laden to September 11 attack
3498	42888217	-	September 11 attack
10117	254875340	<i>per talk, per lead, etc.</i>	editing several parts about linking bin Laden to different terrorist attacks
7594	78510441	-	linking bin Laden to September 11 attack
8725	137684092	-	linking bin Laden to September 11 attack
8731	137795596	<i>erroneous report deleted</i>	linking bin Laden to September 11 attack
7840	84882264	<i>rv, the intro is not for making a case against him</i>	removing evidences about linking bin Laden to different terrorist attacks
10377*	309653330	<i>Osama</i>	changing occurrences of bin Laden to Osama (mentioning only his first name)
1145*	13375039	-	vandalism containing several vulgar words
7902	89767092	<i>I hope it won't be offending, branding him as a wanted fugitive being the first thing people see when they come to this page, but he fits the criteria for inclusion and the old one didn't have much</i>	adding information about bin Laden's terrorist activities to the infobox

Table 7.4: Top-10 selected revisions for Osama bin Laden article

of most controversy-contributing revisions in each controversial article.

We modeled contribution of revisions based on the well-known problem of maximum coverage, where we adapted it to our problem in a way to select the revisions that cover most of the controversial ideas debated in controversial articles. We evaluated the quality of revisions selected by our method at two levels of individual and set levels. The results of these two evaluation schemes suggested that our method is able to select revisions that are more likely to have contributed to controversy of their articles compared to baselines.

Despite these experimental results, it should be noted that specifying which par-

ticular set of revisions were most responsible for making an article controversial is a somewhat *subjective* task. In particular, controversy in Wikipedia articles is a result series of edits and discussions among some editors on one or more disputed issues. Quantifying the effect of these different edits and pinpointing a limited set of them, specially for small sets, can be challenging. Hence, the revisions selected by our method for each controversial article might not be *the most* contributing revisions in the eyes of different users. However, based on the results of the quantitative evaluations provided in this chapter, at least, we can conclude that these selected revisions are more likely to have contributed to overall controversy of their articles compared to other baselines. They also help extracting some of the disputed issues out of thousands of revisions in controversial articles as shown in our qualitative evaluation, and therefore are useful in giving a more fine-grained view.

Finally, the revision-level analysis and our proposed revision-selection method are just our first attempts in providing a fine-grained analysis of controversy in Wikipedia. Future work can focus on better ways for achieving such a goal. For instance, revisions of each article can be divided to sets of related revisions based on factors such as time proximity or topic relatedness. Then, our revision-level analysis can be modified to select the set of sets of related revisions that had the most contribution in making an article controversial, instead of selecting individual revisions. This allows users to better understand specific disputed issues of each controversial article compared to single, unrelated revisions that can cover different issues. Moreover, more work is needed to provide a summary of the selected revisions to help users grasp the main content of these revisions in a better and faster way. Such a summary can be as simple as tag clouds of edited terms weighted by their global and local scores, or can be more advanced to include only a set of important disputed statements.

# Chapter 8

## Conclusion and Future Work

### 8.1 Conclusion

Wikipedia is probably the most commonly-used knowledge source nowadays. Identifying controversial and disputed content in this knowledge-base can benefit many of its readers and editors, as several of its popular and highly-visited articles are among those that have been tagged as being controversial. Having an accurate, and efficient controversy model can be beneficial to Wikipedia's community as it can help enhancing the current manual process of labeling controversial articles, freeing editors and admins to focus on more creative tasks. It can also be useful in creating a knowledge-base of known controversial topics or can be used as training labels in determining controversy degree of other materials on the web as done in a recent work [17].

The problem of identifying controversial articles can be overlooked, assuming that simple heuristics such as the number of revisions of an article, or the length of its discussion page are enough to identify these articles. However, as shown in this thesis, not only these statistics are not effective alone, but several previous controversy models are inefficient in fully capturing the complex process of formation of controversy in Wikipedia articles.

Collaboration networks are effective ways for characterizing Wikipedia articles. These networks are an abstract representation of collaboration history of each article containing its main contributors connected with signed edges, denoting attitude of editors. In this thesis, we used these networks to identify controversial articles

by extracting a set of structural features and training the Structural classifier over labelled controversial and non-controversial articles.

As discussed, the success of collaboration networks in identifying controversial articles highly depends on the model used to infer attitudes of editors and assign signs of edges. Even though editors do not directly express their attitudes towards one another in Wikipedia, there are different resources that reveal the type of interactions and collaboration of editors. Revision history of articles and admin election repository are two of such resources that were exploited in this thesis. The former was used in the form of several global features to build collaboration profiles of editors, while the later was used as training labels for classifying these profiles.

The combination of this attitude inference model and the structural features extracted from the signed networks of collaborations resulted in a highly accurate method for identifying controversial articles as shown by an extensive experimental validation and comparison with other methods. Also the experimental results showed that this controversy model is remarkably accurate when only a fraction of the revision history is available or when working with articles with fairly short histories. These can be attributed to the global features that were collected for editors across the entire Wikipedia, and to the set of structural features extracted from each network, where many of which are rooted at sound theories of social behavior.

In addition, the analysis of controversy can greatly improve the experience of users of Wikipedia if it provides them with main arguments and opposing views causing articles to become controversial. The unit-level analysis, and revision-level analysis are two approaches proposed in this thesis towards this goal, which each focuses on a specific type of article element (i.e. text-units for unit-level analysis, and revisions for revision-level analysis) to locate sources of controversy.

The unit-level analysis turned-out be a challenging problem, mainly due to difficulties in creating a benchmark for evaluating different models. This suggests that the lack of a standard ground truth is the main obstacle in fine-grained analysis of controversy, which requires further attention and work in future.

Compared to unit-level analysis, the revision-level analysis was shown to be a more successful approach for fine-grained analysis, where the well-known frame-

work of maximum coverage problem was used to define contribution of revisions in our Coverage-based Controversy Contribution (CCC) method. The experimental results demonstrated the effectiveness of CCC as a contribution model that naturally satisfies the desired computational properties due to fitting in maximum coverage framework. This model is also a general and intuitive model that can be adjusted depending on different contexts and needs. For instance, term-global and term-local scores were defined in this thesis based on different heuristics found to be useful in detecting controversy in Wikipedia articles. However, one can think of other effective heuristics for this medium or other media and easily incorporate them in the definition of these two scores to come up with other contribution models.

## **8.2 Future Works**

### **8.2.1 Improving Attitude Inference Model**

As we showed, the performance of our controversy model highly depends on the method used to infer the attitudes of editors and assigning the signs of edges in collaboration networks. In lack of ground truth, we used admin elections to indirectly learn these attitudes. However, despite its effectiveness, such a source is specific to Wikipedia and is not available in most other wikis and collaborative systems. Hence, future work might try to develop other ways for this purpose based on other available evidences such as analysis of discussion pages, or comments of revisions. Moreover, we considered a single, global attitude for each pair of interacting editors, due to using admin elections to infer attitudes of editors. If other effective ways are developed in future to infer attitudes of editors, one might be able to infer attitudes of editors at more fined-grained levels, such as at article, or category levels. It is interesting to assess the performance of these fined-grained attitudes based on the evaluation methods we used in this thesis, and see how they perform in comparison with our current global-level approach.



## 8.2.2 Modeling Controversy in other Domains

From a more general perspective, our notion of collaboration networks and the Structure classifier can be applied to other user-driven media, such as discussion forums and comments made to blogs and news posts. For instance, there are many attempts on identifying agreement/disagreement relations between participants in meetings [23, 28, 74] or building a signed network of users for discussion forums [27, 51, 56, 71]. While different approaches have been proposed to identify these relations or build these networks, the methods are either tested on topics *known* to be controversial, or no controversy analysis has been done. Thus, we posit that applying Structure classifier on the extracted relations and built signed networks would help detect whether the underlying discussion is controversial, or not.

## 8.2.3 Representation of Selected Revisions

The revision-level analysis we conducted helps users to have better insights about the disputed issues in controversial articles by separating normal, and vandalism revisions from those that contributed to controversy and get a ranked list of these revisions. However, the user still has to manually investigate these revisions and the changes that have been applied in them to be able to grasp a knowledge about these issues. What can be more useful is to represent these revisions in a more abstract way, summarizing the main disputed issues. This summarization in its simplest form can be done by listing top selected terms based on local and global scores of terms in our coverage-based contribution model. At a more advanced level, summarization techniques that contrast multiple documents and summarize their similarities and differences [52, 77, 78] can be used. Alternatively, terms in our coverage-based approach can be defined based on higher-level concepts such as topics in topic modeling approaches, which allows to be able to represent the final selected revisions at a more abstract way. Moreover, the revision-level analysis can be improved to select set of sets of related revisions instead of set of individual and unrelated revisions, where relatedness of revisions can be defined based on factors

such as time proximity or topic closeness. In this way, users will be represented with sets of related revisions, where each set is likely to be about a different disputed issue, and thereby easier for users to interpret.

#### **8.2.4 Considering Other Factors in Selecting Revisions**

Future work can improve the revision-level analysis by considering factors other than the individual contribution of revisions. For instance, we might request to select a diverse set of contributing revisions that spans different issues instead of selecting all contributing revisions. As another example, the duration and importance of the underlying disputed issue that each revision belongs to might be considered in selecting contributing revisions. In this way, the objective function  $F$  introduced in Equation 6.1 should be designed in a way to take these factors into account, while satisfying the desired computational properties.

#### **8.2.5 Other approaches for fine-grained Analysis of Controversy**

Both unit-level and revision-level analyses mainly rely on the revision history of articles and the meta data they contain. Hence, what they can extract is limited to disagreements and disputes that are explicitly expressed through different edit actions, and does not include the opposing views and conflicts discussed in the discussion pages. These discussions contain more detailed information and reasoning about the disputed issues compared to changes recorded in the revision history. Hence, other approaches for fine-grained controversy analysis can include the ones that utilize these valuable resources and analyze them by means of natural language processing techniques. With recent attempts on annotating discussion pages, and more advancement on analyzing discussion forums [4, 16, 51, 64] there is more hope for developing such methods in future.

# Bibliography

- [1] B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. Assigning trust to Wikipedia content. In *Proceedings of the 4th International Symposium on Wikis, WikiSym '08*, pages 26:1–26:12, New York, NY, USA, 2008. ACM.
- [2] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 261–270, New York, NY, USA, 2007. ACM.
- [3] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. Harmony and dissonance: Organizing the people’s voices on political controversies. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12*, pages 523–532, New York, NY, USA, 2012. ACM.
- [4] Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media, LSM '11*, pages 48–57, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [5] Petko Bogdanov, Nicholas D. Larusso, and Ambuj Singh. Towards community discovery in signed collaborative interaction networks. In *Proceedings of the IEEE International Conference on Data Mining Workshops, ICDMW '10*, pages 288–295, Washington, DC, USA, 2010. IEEE Computer Society.
- [6] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. Societal controversies in Wikipedia articles. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15*, pages 193–196, New York, NY, USA, 2015. ACM.
- [7] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proceedings 1st International Workshop Motivation and Incentives on the web, Webcentives '09*, pages 4–11. ACM, 2009.
- [8] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, pages 731–740, New York, NY, USA, 2009. ACM.
- [9] Ulrik Brandes and Jürgen Lerner. Visual analysis of controversy in user-generated encyclopedias. *Information Visualization*, 7(1):34–48, March 2008.

- [10] Sjarhei Bykau, Flip Korn, Divesh Srivastava, and Yannis Velegrakis. Fine-grained controversy detection in Wikipedia. In *Proceedings of 31st IEEE International Conference on Data Engineering, ICDE'15*, pages 1573–1584, 2015.
- [11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web, WWW '11*, pages 675–684, New York, NY, USA, 2011. ACM.
- [12] Krishnendu Chatterjee, Luca de Alfaro, and Ian Pye. Robust content-driven reputation. In *Proceedings of the 1st ACM Workshop on Workshop on AISEc, AISEc '08*, pages 33–42, New York, NY, USA, 2008. ACM.
- [13] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, pages 355–362, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [14] Naeemah Clark. Trust Me! Wikipedia's Credibility Among College Students. *International Journal of Instructional Media*, 38(1):27–36, 2011.
- [15] Johannes Daxenberger and Iryna Gurevych. Automatically classifying edit categories in Wikipedia revisions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 578–589, Seattle, WA, USA, 2013. Association for Computational Linguistics.
- [16] Laura Dietz, Ziqi Wang, Samuel Huston, and W. Bruce Croft. Retrieving opinions from discussion forums. In *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management, CIKM '13*, pages 1225–1228, New York, NY, USA, 2013. ACM.
- [17] Shiri Dori-Hacohen and James Allan. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management, CIKM '13*, pages 1845–1848, New York, NY, USA, 2013. ACM.
- [18] Shiri Dori-Hacohen, Elad Yom-Tov, and James Allan. Navigating controversy as a complex search task. In *Proceedings of the 1st CEUR International Workshop on Supporting Complex Search Tasks*, 2015. Electronic proceedings only.
- [19] Gregory Druck, Gerome Miklau, and Andrew McCallum. Learning to predict the quality of contributions to Wikipedia. In *Proceedings of 1st International AAAI Workshop on Wikipedia and Artificial Intelligence, WAI*, pages 7–12. AAAI Press, 2008.
- [20] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12*, pages 63–72, New York, NY, USA, 2012. ACM.
- [21] Katherine Faust. A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*, 32(3):221 – 233, 2010.

- [22] Fabian Flöck, Denny Vrandečić, and Elena Simperl. Towards a diversity-minded Wikipedia. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 5:1–5:8, New York, NY, USA, 2011. ACM.
- [23] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [24] Mark Granovetter. The strength of weak ties: A network theory revisited. In P. V. Marsden and N. Lin, editors, *Social Structure and Network Analysis*, pages 105–130, Beverly Hills, CA, 1982. Sage Publications.
- [25] Karim Hajian-Tilaki, James Hanley, Lawrence Joseph, and Jean-Paul Collet. A comparison of parametric and non-parametric approaches to roc analysis of quantitative diagnostic tests. *Medical Decision Making*, 17(1):94–102, 1997.
- [26] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Automatic quality assessment of content created collaboratively by web communities: A case study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 295–304, New York, NY, USA, 2009. ACM.
- [27] Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 59–70, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [28] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short '03, pages 34–36, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
- [29] <http://en.wikipedia.org/wiki/Wikipedia>. Wikipedia. Last visited on July 9, 2015.
- [30] [https://en.wikipedia.org/wiki/Wikipedia:Systemic\\_bias](https://en.wikipedia.org/wiki/Wikipedia:Systemic_bias). Wikipedia:systemic bias. last visited on september 26, 2015.
- [31] <http://stats.grok.se/en/top>. Wikipedia article traffic statistics. last visited on july 9, 2015.
- [32] <http://www.nytimes.com/2011/01/31/business/media/31link.html>. Define gender gap? look up Wikipedias contributor list. last visited on september 26, 2015.
- [33] <http://www.nytimes.com/2014/02/10/technology/wikipedia-vs-the-small-screen.html>. Wikipedia vs. the Small Screen. Last visited on July 9, 2015.

- [34] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in Wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07*, pages 243–252, New York, NY, USA, 2007. ACM.
- [35] J. W. Hunt and M. D. McIlroy. An algorithm for differential file comparison. Technical Report CSTR 41, Bell Laboratories, Murray Hill, NJ, 1976.
- [36] Sara Javanmardi, Cristina Lopes, and Pierre Baldi. Modeling user reputation in wikis. *Statistical Analysis and Data Mining*, 3(2):126–139, 2010.
- [37] Sara Javanmardi, David W. McDonald, and Cristina Videira Lopes. Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym'11*, pages 82–90, 2011.
- [38] Taghi M. Khoshgoftaar, Moiz Golawala, and Jason Van Hulse. An empirical study of learning from imbalanced data using random forest. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02, ICTAI '07*, pages 310–317, Washington, DC, USA, 2007. IEEE Computer Society.
- [39] Aniket Kittur, Bongwon Suh, and Ed H. Chi. Can you ever trust a wiki?: Impacting perceived trustworthiness in Wikipedia. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, CSCW '08*, pages 477–480, New York, NY, USA, 2008. ACM.
- [40] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07*, pages 453–462, New York, NY, USA, 2007. ACM.
- [41] Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 Conference Short Papers, ACLShort '10*, pages 253–257, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [42] Thomas Rune Korsgaard and Christian D. Jensen. Reengineering the Wikipedia for reputation. *Electronic Notes in Theoretical Computer Science.*, 244:81–94, August 2009.
- [43] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 641–650, New York, NY, USA, 2010. ACM.
- [44] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07*, pages 420–429, New York, NY, USA, 2007. ACM.
- [45] Chenliang Li, Anwitaman Datta, and Aixin Sun. Mining latent relations in peer-production environments: a case study with Wikipedia article similarity and controversy. *Social Network Analysis and Mining*, 2(3):265–278, 2012.

- [46] Hui Lin and Jeff Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Uncertainty in Artificial Intelligence (UAI)*, pages 479–490, Catalina Island, USA, 2012. AUAI.
- [47] Wei-Hao Lin and Alexander Hauptmann. Are these documents written from different perspectives?: A test of different perspectives based on statistical distribution divergence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 1057–1064, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [48] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 109–116, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [49] Nedim Lipka and Benno Stein. Identifying featured articles in Wikipedia: Writing style matters. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1147–1148, New York, NY, USA, 2010. ACM.
- [50] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.
- [51] Yue Lu, Hongning Wang, ChengXiang Zhai, and Dan Roth. Unsupervised discovery of opposing opinion networks from forum discussions. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1642–1646, New York, NY, USA, 2012. ACM.
- [52] Inderjeet Mani and Eric Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2):35–67, May 1999.
- [53] Silviu Maniu, Bogdan Cautis, and Talel Abdesslem. Building a signed network from interactions in Wikipedia. In *Databases and Social Networks*, DB-Social '11, pages 19–24, New York, NY, USA, 2011. ACM.
- [54] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- [55] Santiago M. Mola-Velasco. Wikipedia vandalism detection. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 391–396, New York, NY, USA, 2011. ACM.
- [56] Akiko Murakami and Rudy Raymond. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd international conference on Computational Linguistics: Posters*, COLING '10, pages 869–875, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [57] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

- [58] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008.
- [59] Souneil Park, KyungSoon Lee, and Junehwa Song. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 340–349, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [60] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1873–1876, New York, NY, USA, 2010. ACM.
- [61] Martin Potthast and Teresa Holfeld. Overview of the 2nd international competition on Wikipedia vandalism detection. In *CLEF 2011 Labs and Workshop, Notebook Papers*, 2011.
- [62] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, GROUP '07, pages 259–268, New York, NY, USA, 2007. ACM.
- [63] David M. Reif, Alison A. Motsinger, and Brett A. McKinney. Feature selection using a random forests classifier for the integrated analysis of multiple data types. In *proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–8. IEEE, 2006.
- [64] Jodi Schneider, Alexandre Passant, and John Breslin. A qualitative and quantitative analysis of how Wikipedia talk pages are used. In *proceedings of the 2nd international conference on WebScience*, pages 1–7. ACM, 2010.
- [65] Hoda Sepehri Rad and Denilson Barbosa. Towards identifying arguments in Wikipedia pages. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 117–118, New York, NY, USA, 2011. ACM.
- [66] Hoda Sepehri Rad and Denilson Barbosa. Identifying controversial articles in Wikipedia: A comparative study. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 7:1–7:10, New York, NY, USA, 2012. ACM.
- [67] Hoda Sepehri Rad and Denilson Barbosa. Identifying controversial Wikipedia articles using editor collaboration networks. *ACM Transaction on Intelligent Systems and Technology (TIST)*, 6(1):1–24, March 2015.
- [68] Hoda Sepehri Rad, Aibek Makazhanov, Davood Rafiei, and Denilson Barbosa. Leveraging editor collaboration patterns in Wikipedia. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 13–22, New York, NY, USA, 2012. ACM.
- [69] Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational*



- Linguistics*, EACL '12, pages 224–233, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [70] Ruben Sipos, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. Temporal corpus summarization using submodular word coverage. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 754–763, New York, NY, USA, 2012. ACM.
- [71] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 226–234, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [72] Bongwon Suh, Ed H. Chi, Bryan A. Pendleton, and Aniket Kittur. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, VAST '07, pages 163–170, Washington, DC, USA, 2007. IEEE Computer Society.
- [73] Hiroya Takamura and Manabu Okumura. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 781–789, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [74] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 327–335, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [75] Tadej Štajner, Bart Thomee, Ana-Maria Popescu, Marco Pennacchiotti, and Alejandro Jaimes. Automatic selection of social media responses to news. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 50–58, New York, NY, USA, 2013. ACM.
- [76] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw, and Kuiyu Chang. On ranking controversies in Wikipedia: Models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 171–182, New York, NY, USA, 2008. ACM.
- [77] Xiaojun Wan, Houping Jia, Shanshan Huang, and Jianguo Xiao. Summarizing the differences in multilingual news. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 735–744, New York, NY, USA, 2011. ACM.
- [78] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. Comparative document summarization via discriminative sentence selection. *ACM Transaction on Knowledge Discovery Data*, 6(3):12:1–12:18, October 2012.

- [79] Andrew G. West, Sampath Kannan, and Insup Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata? In *Proceedings of the 3rd European Workshop on System Security, EUROSEC '10*, pages 22–28, New York, NY, USA, 2010. ACM.
- [80] Laurence Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, pages 385–393, 1982.
- [81] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. Characterizing Wikipedia pages using edit network motif profiles. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, SMUC '11*, pages 45–52, New York, NY, USA, 2011. ACM.
- [82] Taha Yasseri, Róbert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in Wikipedia. *PloS one*, 7(6):e38869, 2012.
- [83] Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services, PST '06*, pages 8:1–8:1, New York, NY, USA, 2006. ACM.

# Appendix A

## Appendix

### A.1 Unit-level Analysis assuming Independent Units

When units are assumed to be independent, solving the unit-level analysis is equivalent to ranking units in terms of their individual contribution score, and choosing the top  $k$  units.

**Proof:** With the assumption of independence of contribution of units, for any set  $S$  of units, we have  $F(S) = \sum_{u \in S} f(u)$ . With this assumption, it can be shown that  $S^{rank}$  obtained by choosing the top  $k$  units is the set maximizing the optimization equation 6.1, and hence is equivalent to  $S^*$ . For showing this equivalence, we resort to counter-proof strategy. Let's assume that there is a set  $S'$  whose  $F$  is larger than  $F(S^{rank})$ . First, we sort the units of  $S'$  in terms of their  $f(u)$  in decreasing order. Then, comparing each unit  $u'_i \in S'$  at each position  $i = \{1..k\}$  with the unit  $u_i \in S^{rank}$  in the same position  $i$ , we have  $f(u'_i) \leq f(u_i)$ . This is because each  $u_i$  according to the ranking of units is the largest  $i$ th unit among all units  $U$ , and the  $i$ th largest contributing unit of any set  $S$ , including  $S'$ , cannot have a contribution value larger than this unit. Finally, as  $F(S') = \sum_{u' \in S'} f(u')$ ,  $F(S')$  cannot also be larger than  $F(S^{rank})$ , and thereby  $F(S^{rank})$  is equal to  $F(S^*)$ . Hence,  $S^{rank}$  is the same as  $S^*$  (or one of the possible  $S^*$  solutions).

### A.2 Mutual Reverts is a Monotone Score

Recall from Chapter 4 that Mutual Reverts score for article  $a$  is calculated as follows:

$$MR^a = E^a \times \sum_{N_i^a, N_j^a < max} \min(N_i^a, N_j^a) \quad (A.1)$$

In this equation, the sum is over all editor pairs of  $i$  and  $j$ , who mutually reverted each other's edits at least once. Let us name the set of such pairs as  $mre_a$ . The number of these pairs is shown by the quantity  $E^a$  in the above equation.

Also, as discussed in Chapter 6, a function  $f$  is monotone if:  $\forall A \subset B, F(A) \leq F(B)$ . To verify the monotonicity condition for  $MR$  function, we need to calculate  $MR$  based on a subset of article elements instead of the whole article as in  $MR^a$ . In Chapter 6 and 7, we discussed two different types for these article elements: text-units as in the text-unit level analysis, and revisions as in the revision-level analysis. As discussed in those chapters, both of these types of elements can be considered to be like a synthetic article that contains only edits corresponding to the considered elements. Hence, similar to  $MR^a$  which is calculated based on the whole revision history of article  $a$ ,  $MR^A$  and  $MR^B$  are calculated based on the set of revisions corresponding to sets  $A$  and  $B$  respectively.

Now, as  $A \subset B$ , we have:  $mre_A \subset mre_B$ , which means that  $B$  contains all the pairs of mutually reverting editors of  $A$ , in addition to have zero or more of such pairs. This enforces to have:  $E_A \leq E_B$ . Similarly, for the other term in Equation A.1, we will have:

$$\sum_{N_i^A, N_j^A < max} \min(N_i^A, N_j^A) \leq \sum_{N_i^B, N_j^B < max} \min(N_i^B, N_j^B)$$

Therefore,  $MR^A \leq MR^B$ , and  $MR$  is a monotone function.