PROBABILISTIC MODELS FOR IDENTIFYING AND EXPLAINING CONTROVERSY

A Thesis Presented

by

MYUNGHA JANG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2019

College of Information and Computer Sciences

PROBABILISTIC MODELS FOR IDENTIFYING AND EXPLAINING CONTROVERSY

A Thesis Presented

by

MYUNGHA JANG

Approved as to style and content by:

James Allan, Chair

W. Bruce Croft, Member

Brendan O'Connor, Member

Weiai Wayne Xu, Member

James Allan, Chair of the Faculty
College of Information and Computer Sciences

DEDICATION

To my family and friends

ACKNOWLEDGMENTS

I was very lucky to have found James as my advisor. I have learned a lot from him not just about IR and how to research, but he was a great professional role model to learn from about communication skills, attention to details, patience, and work ethics. As one of the most reasonable people I've ever met, I have always felt at ease for communicating about any issue with him, and it has made my graduate life a lot more stress-free than it could have been. He also has given me a great deal of freedom in choosing a research topic and was patient until I make it through, but he also knew well when to push me when I needed to be pushed.

His humors have lightened up many moments, even during my own crisis. In my first year, I panicked when I realized that there was a small bug in my experiment. The accuracy of my approach that I was going to report in a conference paper that was due soon turned out to be slightly lower than what I originally had shown to James. I ran to James and confessed, "`I'm sorry, I found a small bug in my experiment.`". Sensing my desperate tone, he joked, "`Okay, did you sort the results in a reverse order? Is your method now the worst?`" The joke caught me off guard and lightened me up so much that I felt more comfortable telling him about my mistake. If I can keep going with my favorite anecdotes with James' jokes, I probably need another chapter in this thesis.

I would like to thank my committee members: Bruce Croft, Brendan O'Connor, and Wayne Xu. Brendan O'Connor deserves special credit for sharing his Twitter data, which enabled us a lot of interesting experiments in Chapter 7.

I would like to thank Kenneth Church, who was my internship mentor at IBM Research. He has taught me many research principles and skills that stay relevant with me to this day.

CIIR has been a great lab where I can always learn something from many colleagues: John Foley, Youngwoo Kim, Hamed Zamani, Hamed Rezanejad (Rab), Shahrzad Naseri, Helia Hashmi, Qingyao Ai, Keping Bi, Liu Yang, Ali Montazer, Chen Qu, Sheikh Sarwar, Jiepu Jiang, Dan Cohen, Lakshmi Vikraman, and Yen-Chieh Lien. Especially, "The Night Watch" in the lab, Hamed, Youngwoo, Helia, Sheikh, and Rab made those long nights in the lab feel much less lonely. Youngwoo deserves immense credit for his contributions that help me finish this thesis; for being a great collaborator, a wise friend when I'm being indecisive, a lab Batista , and for pulling my arm when I procrastinate from writing my thesis. I also received a fair amount of mentoring from CIIR alumni in my early years including from Jeff Dalton, Laura Dietz, Weize Kong, Youngho Kim, Ethem Can, Elif Aktolga, Zeki Yalniz, Shiri Dori-Hacohen and Marc-Allen Cartright.

CIIR has awesome staff members that have helped my life significantly easier. I'd like to thank Jean Joyce, Kate Moruzzi, Joyce Mazeski, Victoria Rupp, Glenn Stowell, and Dan Parker for their support. They really got my back from the day one I joined the lab to the the day I graduated.

Leeanne Leclerc, as a former Graduate Program Manager as well as a caring friend, has looked out for us for many years and continued to have our back even after she left the job. When she left the program, it felt like the end of an era, and I was glad that I lived in that era.

I made a lot of great friends who made Amherst feel like home, including Matteo Brucato, Kyle Wray, Tiffany Liu, Tamara Rossi Mercanti, Su Lin Blodgett, Luis Pineda, Niri Karina, Samer Nashed, Justin Svegliato, Pinar Ozisik, Emma Strubell, Pat Verga, Kristen Atwood, James Atwood, Dirk Ruiken, Philip Thomas, Sandhya

Saisubramanian, Lakshmi Vikraman and Emma Tosch. Especially, I thank Kyle and Matteo for their great friendship and many fond memories in the Gray st. house where we play video games or chat about our work, future, relationships, or just about anything, which often went until late at night. Matteo is probably one of the wittiest people I have ever met and I thank him for enriching my life with countless laughs, coffee lessons, and for getting me into Rocket League.

I'd like to thank my 608 housemates, Kevin Winner, Amanda Gentzel, Keen Sung, Li Yang Ku, Larkin Flodin, Joie Wu for many great years in the house.

Of course, I would not have gone through this journey without my Korean friends. Having met my best friend, Sangshin Park, was one of the luckiest thing that has happened to me in Amherst. His emotional support has never failed to hold me up. My another best friend Saejung Kim, the kind of friend who flew from Korea to Amherst to come see me, has been there for me day and night whenever I need to talk. With my fellow Korean CS grad friends, Souyoung Jin, Shinyoung Cho, Yeonsup Lim, and Youngwoo Kim, we have made an awesome community together. They have been there for me for any small milestone and achievement throughout the journey. I was also lucky to have found friends outside of CS: Soo oak Yoo, Kwang-won Park, Dasol Kim, Jaeyoung Ahn, Jaieun Kim, Yunah Kim, Sunmi Kang. Especially, I thank my unofficial Amherst family, Soo Oak and Kwang-won, for taking care of me when I was sick, listening to me when I was sad, and supporting me by training together for 5Ks when I decided to start running.

Last but not least, my real family has been nothing but supportive from the beginning when I decided to move 6,000 miles away from home to pursue a PhD. They have always encouraged me, supported me emotionally and financially, and have sent me many care packages. None of these would have been possible without their support and love.

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Air Force Research Laboratory (AFRL) and IARPA under contract #FA8650-17-C-9118 under subcontract #14775 from Raytheon BBN Technologies Corporation, in part by NSF grant #IIS-0910884, in part by NSF grant number 1819477, and in part by NSF grant #IIS-1217281. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

ABSTRACT

PROBABILISTIC MODELS FOR IDENTIFYING AND EXPLAINING CONTROVERSY

MAY 2019

MYUNGHA JANG

B.S., EWHA WOMANS UNIVERSITY

M.S., POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor James Allan

Navigating controversial topics on the Web encourages social awareness, supports civil discourse, and promotes critical literacy. While search of controversial topics particularly requires users to use their critical literacy skills on the content, educating people to be more critical readers is known to be a complex and long-term process. Therefore, we are in need of search engines that are equipped with techniques to help users to understand controversial topics by identifying them and explaining why they are controversial. A few approaches for identifying controversy have worked reasonably well in practice, but they are narrow in scope and exhibit limited performance.

In this thesis, we first focus on understanding the theoretical grounding of the state-of-the-art algorithm. We derive an underlying probabilistic model that explains the state-of-the-art controversy detection algorithm. We revisit the properties and assumptions from the derived model, and propose new methods to identify controversy

on Webpages. We then point out that the current approaches for controversy detection do not consider *time* while controversy is a dynamically changing phenomenon. This causes current methods to have delays in recognizing emerging controversial topics or exaggerated effects on outdated controversies. We address time-adaptable controversy detection by estimating the dynamically-changing controversy trend of topic by interpolating the observed level of contention and the public interest over time on the topic. Finally, we offer a method that explains controversy by generating a summary of each stance. Our method ranks social media postings using a score of how likely it is that the given post can be a representative summary of controversy.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

As the primary sources for information are now online (Mitchell et al., 2016), the internet and social media have a bigger influence than ever on people's decisions across various domains of real-life problems. While the information that people access might have a tangible and beneficial impact on decisions they make, there is a caveat: people are easily exposed to lots of biased, unscientific, unproven, untrustworthy, or fake information, which reflects that the topic being researched might be controversial. For this reason, search of controversial topics in particular requires users to be extra careful not to be misled. In addition, there are a few other factors that cause understanding the search results of controversial topics to be more chaotic and challenging. As some controversial topics tend to change quickly, the amount of information needed to catch up quickly grows to be overwhelming for users. To make it worse, while social media is one popular place where controversial discourse is held, its "echo chamber" phenomenon limits users from accessing diverse perspectives on controversial topics.

To set the stage, we first discuss these factors that make search of controversial topics particularly challenging. We then briefly discuss the philosophical question raised around the "facilitator" role that a search engine is expected to play in promoting critical literacy. We argue the necessity of a controversy-aware search system as a solution to help users to navigate controversial topics and introduce our technical contributions and challenges towards that goal.

1.1 Challenges of Search for Controversial Topics

There are a few factors that make search of controversial topics a particularly challenging task. We discuss three aspects here: misinformation, information overload, and the echo chamber phenomenon.

1.1.1 Misinformation on the Web

As anyone is free to publish anything on the internet, misinformation or unverified information is prevalent on the Web. Medicine is one of the fields that frequently faces challenges with misinformation, for example, fraudulent treatments or spurious links between two factors such as vaccines and autism. In fact, a recent study shows that misinformation contained in search results and spread through the social network threatens public health (Vogel, 2017). Vaccination is one good example of this issue. In 2014, the United States had one of the largest recent measles outbreaks, which was caused by vaccine hesitancy (Pannaraj, 2018). Brunson et al. (2013) studied the impact of the social network on parents' vaccination decisions for their children. The study found that parents rarely make their decision alone on whether their child should be vaccinated or not, but resort to online sources to find information and advice before making a decision. The influence of the social network was huge particularly for parents who do not vaccinate at all.

Information that users are exposed to in the political sphere also has a significant influence on people's decisions and votes. For example, users might search for presidential candidates to learn about their campaigns or last night's presidential debate to make up their mind for whom to vote during a presidential election. Allcott and Gentzkow (2017) explained that the evidence suggests that false information (or "fake news") spread throughout the social network might have changed the result of 2016 U.S. Presidential election. The evidence includes that (1) 62% of U.S. adults use social media as their primary source of news (Gottfried and Shearer, 2016), (2) the

most popular fake news stories went more "viral" than the real news on Facebook (Silverman, 2016), (3) 75% of American adults who saw fake news headlines viewed them as accurate (Silverman and Singer-Vine, 2016), and (4) the most popular fake news stories tended to be in favor of Donald Trump over Hillary Clinton (Silverman, 2016). After the election, several commentators analyzed the situation and ended up suggesting that Donald Trump would not have been elected without the influence of fake news on Facebook (Parkinson, 2016; Read, 2016; Dewey, 2016). However, the study by Guess et al. (2018) also suggests that most fake news were consumed by Trump supporters. Whether or not the result of the election would have changed, this demonstrates how significant the effects of misinformation can be to our society, especially for high-stake controversial topics.

1.1.2 Information Overload

Shahaf and Guestrin (2010) discuss the information overload problem wherein despite extensive media coverage, people often have difficulty understanding a news event. For example, David Leonhardt's New York Times article, "Can't Grasp Credit Crisis? Join the Club" suggests that while many people probably felt as if they should understand the credit crisis with so many stories published, many of them actually didn't understand (Leonhardt, 2008). Because the amount of information on a controversial topic quickly grows to be huge, especially when controversy develops from a "scandal" into a "saga" (Cramer, 2011), it is difficult to stay up-to-date while controversy is happening if you are not closely following the case. Therefore, addressing the information overload problem to help people understand a controversial topic is another critical issue that we need to deal with.

To address this, creating a summary of events in a chronological order has been studied as a solution to help users understand a dynamically-changing news event (Shahaf and Guestrin, 2010; Allan et al., 2001). However, existing techniques do not

focus on understanding the aspects of controversy within the event. Therefore, an algorithmic solution that explains the event from a controversial perspective is needed to directly handle questions such as "why is this case controversial?" and "what are the conflicting stances and discourses that are being discussed around this controversy?".

1.1.3 Echo Chambers

Social media's news feed algorithms are intentionally biased toward connecting like-minded people, assuming that users would like to see information they are likely to agree with. Such algorithmic bias and the growing polarization on controversial topics have resulted in and contributed to the spread of a "filter bubble" or "echo chambers" where users are segregated from other viewpoints that are different from their own (Pariser, 2011; Jackson, 2017). For example, for users who search for a controversial topic on social media to understand what is going on, current search system makes this navigation difficult as the top posts are likely to be the ones that the user agrees with because her friends "liked" the posts or she or her friends follow the authors. This prevents users from obtaining a balanced holistic view of the issue. As users get more exposed to content tailored to their view, this echo chambers phenomenon strengthens over time, causing a vicious cycle (Garimella, 2018).

1.2 The Role of Search Engine to Promote Critical Literacy

Critical literacy (Wikipedia, 2019a) is the ability to identify possible bias or discrimination that the author might have projected in her writing. In an ideal world, users are well-equipped with critical literacy skills and actively practice them when they read documents on the Web. In reality, people are more likely to be trusting, especially when they are not even aware that the topics that they are searching for are controversial. Educating people to be more critical readers is a complex and long-term process (Lapowsky, 2017). In the United Kingdom, while the national

curriculum includes critical literacy skills in every stage, surveys show that 20% of students tend to believe everything that they read on the internet and 30% of UK teachers say that students have cited false information found on the internet for their assignment (Douglas, 2017).

Whether or not, and how a search system should be involved to address the issues mentioned above are a rather philosophical questions While some believe that the spread of misinformation on the Web should be blocked by identifying fake news, others feel repulsed by the idea of censorship, and are not interested in being told that something is not correct when they feel that it is true (Kolbert, 2017). While Garimella at al. (2017) proposed an algorithmic solution to reduce controversy by connecting people with opposing views on social media, some argue that people do not actually want to get out of their echo chambers (Wiseman, 2016). Some experts believe that technical solutions will not decrease the spread of misinformation because technology will create more challenges that will not be countered at scale. A counter argument is that technology will help label, filter, or ban misinformation and aid people to be more critical readers (Anderson and Raine, 2017).

While how much a search engine should "meddle" as a facilitator for controversial topics is left as a controversial issue itself, we argue that a system should at least be "aware" of controversial topics and assist users to navigate controversial topics more effectively by addressing misinformation on the Web, information overload, and echo chambers, to promote critical literacy. Doing so allows the system to act as a minimal facilitator at least by attempting to provide meta-information to give users sufficient perception to decide what to trust and what not to trust, explore other opinions, and understand the various stances of controversial topics. Hence, we propose a develop a controversy-aware search system.

We define *controversy-aware search systems* to refer systems that adopts algorithmic solutions to process the search results of controversial topics. The goal of the

system includes not only helping users who actively seek to understand some controversial topic, but also alerting users who are not even aware that this topic that they are reading is controversial. Therefore, the system overall aims to modulate the search results for controversial topics by systematically mediating the bias and the filter bubble phenomenon.

1.3 Contributions

This thesis covers the following three topics:

- Modeling controversy detection in Web documents,
- Estimating temporal controversy score trend, and,
- Explaining controversy on social media

We discuss the technical contributions under each topic.

1.3.1 Modeling Controversy Detection in Web documents:

1.3.1.1 Deriving a probabilistic framework

To understand the model behind the prevailing algorithms for controversy detection, we analyze the state-of-the-art algorithm (kNN-WC) (Dori-Hacohen and Allan, 2015) and derive an underlying model that explains the theoretic grounding of the algorithm. We show that the underlying model has two probability components: the probability that a document d retrieves a Wikipage w as a topic, and the probability that the people in the relevant population (i.e., Wikipage editors) of w are in contention. We identify the following properties that the model holds:

- **P1:** kNN-WC model uses a population-based topic controversy model as a sub-component.

- **P2:** kNN-WC model does not directly model "non-controversiality". While the model is tuned to capture the mention of controversial topics, the model does not actively take into consideration of the balance of the non-controversial content of the document.

- **P3:** The text of a query document is only used as a proxy to retrieve documents' topics and does not directly affect the probability that the document is controversial.

1.3.1.2 Improving the kNN-WC algorithm

- We revisit the kNN-WC algorithm, which is the specific implementation that Dori-Hacohen and Allan proposed, and assess how accurately this algorithm implements the derived model. In order to implement the kNN-WC model accurately, two probabilistic components are expected to properly estimated. We point out that the algorithm often fails to meet these assumptions. We propose two modifications to improve the accuracy of each probability to better implement kNN-WC model. We suggest two solutions to fix the based on the two findings: First, generating multiple queries from several semantically-coherent paragraphs is more effective in finding relevant Wikipedia topics. Second, since a controversial discussion that contributes to a controversy score usually takes place in a few representative pages among Wikipedia pages of similar controversial topics, smoothing the controversy score from taxonomically-related Wikipedia pages makes the controversy score more accurate.

- We evaluate the proposed solutions both intrinsically and extrinsically. To intrinsically evaluate a new query method, TILEQUERY, to find k Wikipages, we curate a new annotated dataset that includes relevance judgments on the Wikipages for the query documents that are used to for controversy detection.

The new algorithm that combined the two fixes significantly improves the controversy detection task in Webpages by 6% (Jang and Allan, 2016).

1.3.1.3 Proposing Controversy Language Model

- We propose an alternative Controversy Language Model (CLM) where all three properties (P1, P2, and P3) are challenged. Instead of having a population-based topic controversy model as a sub-component, which requires the explicit "contention" features, its "contention" feature was transformed to a "language" feature by building a language model from contentious topics (challenging P1). CLM also directly captures the probability that a document is non-controversial by explicitly considering the probability that the document is generated from controversial topics and non-controversial topics (challenging P2). Lastly, CLM directly considers the document's text to estimate the probability of controversiality (challenging P3).

- To evaluate its efficacy, we experiment with various ways of constructing controversial topics. We show that a CLM that is built with Wikipedia articles that contain several controversy-related keywords was 14% more effective in AUC in identifying controversial Webpages in our dataset, significantly outperforming the kNN-WC algorithm (Jang et al., 2016)

- We compare the characteristics of the the kNN-WC model and CLM via a qualitative analysis. We show that the the kNN-WC+ (our improved version) algorithm is slightly more prone to make false negative errors whereas CLM is more prone to make false positive errors. Short documents tend to be classified as controversial by CLM whereas the kNN-WC+ algorithm has the opposite tendency, compared to the human labels. We present a case study to explain the cases where each algorithm makes a classification error.

1.3.2 Predicting Controversy Score Trend over Time

- We focus on the fact that existing topic-controversy models do not take *time* into consideration. As existing Wikipedia controversy models have used accumulated edit history, the controversy scores do not accurately reflect the true level of controversy that changes over time. Therefore, we develop a new controversy function that estimates the controversy score trend over time. We first investigate a straightforward solution of computing the automated controversy scores by only considering the signals that occurred for a window of given time. We show the trend for topics, even for highly controversial topics, to be highly bursty, and zero for the majority of the time except for the bursty regions. We suggest that generating a temporal controversy score by simply considering a time-window usually yields an unrealistic and impractical trend line.

- We argue that the "observed controversy" does not always accurately reflect the "true controversy" and propose to distinguish the two concepts. We propose that "true controversy" can obtained from considering the two factors, the level of contention and the public interests. We introduce three methods that estimate the true controversy trend by interpolating the trend of the observed controversy obtained from M scores and the public interests obtained from Google Trends.

- We provide a qualitative analysis on the predicted trend line of controversy for various topics.

1.3.3 Explaining controversy on Social Media

- We pose the novel problem of explaining controversy on Twitter via generating a summary of two conflicting stances that make up the controversy. We first characterize a few aspects that a desirable summary should satisfy, namely: stance-indication, articulate level, and topic-relevance.

- We hypothesize that hashtags contain useful information for stance identification and investigate the utility of hashtags in the stance detection task. We train tweet embedding using hashtags as labels to obtain the probability that tweets are likely to generate a given hashtag, for all hashtags. We predict the top relevant hashtags to the given tweet and augment the tweet with them. Using a publicly available stance identification tweet dataset, we show that the when predicted hashtags are added to ngrams of the original tweet text as text features, the F1 score of the stance identification increases from 1% to 5% points.

- We propose a ranking model to rank the tweets by how likely they are to become a good summary to explain controversy. It defines good summary tweets as those whose stance is clearly indicated, whose language is articulate, and whose content is relevant to the given controversial topic. Specifically, we use Twitter's retweet network property to first find user stance communities, and extract the stance hashtags that are distinctively used in each community. We show that tweets are semantically-close to the top stance hashtags best describe the stance community. Being articulate and relevant to the topic makes them even more likely to be an effective summary.

- We evaluate the quality of the ranked tweets as a summary using Amazon Mechanical Turk, compared to other summaries generated from baselines including the state-of-the-art tweet summarization technique. Our human evaluation shows that our summaries are preferred over other baseline summaries (Jang and Allan, 2018).

1.4 Challenges

Building controversy-aware search systems is challenging because navigating controversy is a complex search task for a few reasons. One reason is that determining the extent of the role of the search system is a complicated issue. Dori-Hacohen et al. (2015) brought up two open questions that need to be considered regarding the role of the search system. First, how much should the system help users explicitly in finding content of different stances? For example, should the system only show the results that match the keywords of the user queries even if the result contains the biased results, or make users aware that there are other stances if the query involves controversial topics? Second, should the system deliver every result available, even those that are ungrounded, fraudulent, and even harmful? For example, should the system still present a document of "Issel treatment" as a result for "cancer treatment" when the document contains the query if the system knows that it is also listed as a "dubious treatment" by QuackWatch.com[1]?

In addition to these ethical aspects, search of controversial topics bears numerous technical challenges. While the sub-tasks have a different set of specific challenges, the commonly-shared challenge is that there is a multitude of subtleties in information of controversial topics. For example, while some topics might have a single correct answer, others, especially those that require moral judgment, have several possible answers. The same topic can be controversial to those who care more and know more details about it, while it is not controversial to those who either don't care or don't know much about it (Jang et al., 2017). For these reasons, it is even challenging to computationally define controversy, hence making other related tasks (e.g., recognizing controversy, explaining controversy) inherently difficult.

[1]QuackWatch is a website that allows people to report health-related frauds, myths, or any quackery-related information in medicine.

Unfortunately, prevailing techniques in information retrieval, which are typically designed for retrieving *relevant* information, are not optimized for controversy search. For example, existing search engines are unlikely to reveal controversial topics to users unless they already know about them (Gerhart, 2004). There is a higher call for search engines to detect these queries and address them appropriately (Dori-Hacohen et al., 2015). Earlier work presented an algorithm for classifying controversy in Web documents (Dori-Hacohen and Allan, 2015; Jang and Allan, 2016). However, social media is also increasingly a place where controversy discourse is being shaped and dynamically evolves. Regrettably, we currently lack a tool for effectively navigating the postings around controversy in social media. For example, users have to manually examine postings to find the arguments of conflicting stances that make up the controversy.

Towards the goal of building a system that supports controversy-aware search, we investigate approaches to handle two types of questions: (1) "Does this document discuss a controversial topic?" and (2) "Why is this topic controversial?" While the second task is novel as we propose, the first task has been handled via techniques that classify a document whether it discusses a controversial topic. There have been several algorithms that have been targeted for this task (Dori-Hacohen and Allan, 2013; Dori-Hacohen and Allan, 2015; Beelen et al., 2017; Jang and Allan, 2016), however, little work has explored this problem from a modeling perspective. Therefore, gaps still remain in our theoretical and practical understanding. In this thesis, we study probabilistic models that address the above two questions but that also have an explanatory power in them.

1.5 Thesis Organization

The remainder of this thesis is organized in the following manner. In Chapter 2, we introduce the excising work on controversy detection on the Web and stance

summarization on social media. In Chapter 3, we introduce a probabilistic framework for controversy detection on the Web. We point out that while the state-of-the-art algorithm has proven to perform well empirically, its lack of theoretical underpinning leaves a gap for our understanding. By deriving a theoretical model behind the algorithm, we identify two major assumptions that the model is built on and three properties that the model presents. Subsequently, in Chapter 4, we revisit these assumptions and argue that the algorithm makes erroneous predictions mainly when it fails to reflect the assumptions. To address these challenges, we propose an improved version of the state-of-the-art algorithm by developing two solutions that more accurately reflect the assumptions. In Chapter 5, we revisit the two properties identified from the theoretical model and challenge these properties to propose a new model, controversy language model. In Chapter 6, we propose a method that estimates the "true controversy" score trend that changes over time by interpolating the "observed controversy" with public trend. In Chapter 7, we explore a new problem of summarizing controversy on social media and propose a probabilistic model to rank tweets. Finally, in Chapter 8, we summarize our contributions in this thesis and discuss future research directions in this area.

CHAPTER 2

BACKGROUND

This chapter reviews related work that has been done in the area of controversy detection on the Web and explaining controversy on social media. We discuss the tasks and effort to address them that have been proposed by the research community and how our work builds on them.

2.1 Models in Controversy Detection

Detection of controversy has been mostly studied within a specific online medium such as Wikipedia, social media and online news forums. Existing algorithms, depending on the query, can be categorized into two types: a *topic* controversy model and a *document* controversy model.

2.1.1 Topic Controversy Models

Topic controversy models take a topic as a query and determine the probability that a query topic is controversial. While a topic is loosely defined here, it can be defined from an unstructured format such as any keyword to a specific type of knowledge such as Wikipedia articles[1], hashtags in social media, or named entities. There have been two major aspects in terms of research challenges in designing a new topic controversy model. First, it is how to define and capture controversy, a relatively subjective social phenomenon, from a computational perspective. Our

[1]While Wikipedia articles can technically also be viewed as documents, most existing work in controversy detection consider Wikipedia as a knowledgebase and their articles as topics rather than general documents as the documents contain meta-data and auxiliary edit-history information.

work (not part of this thesis) was the first effort that explicitly investigated the formal definition of controversy (Jang et al., 2017), and argued that "contention" among people and "importance" of the topic are at least two primary dimensions that comprise controversy. However, the "importance" in this context was measured by the number of people to whom the topic mattered. Hence, it can also be represented as "popularity" or "public interest" in the topic.

While prior work other than our work had not explicitly discussed the definition of controversy, most prior work seemed to have the notion of "contention" and "popularity" in their mind in designing an algorithm to identify controversy as most work has focused on capturing signs of "disputes" or "conflicts" among people. Another aspect of the research challenge has been how to capture a few major factors that comprise controversy, particularly "contention", which is characterized and hinted at in a different way in each medium.

For example, in Wikipedia, editors could revert others' changes back and forth when they disagree with each other (Yasseri et al., 2012), whereas in Twitter, users argue back and forth in a thread or exclusively endorse opinions of those who hold the same view as theirs. Such user behaviors can be captured by analyzing a network structure, such as the connectivity between identified retweet communities (Awadallah et al., 2012; Garimella et al., 2016) or motifs of local user interaction (Coletto et al., 2017). Because existing work utilizes the signals that are generated by people who engage in conflicts and disputes, we call this type of model a "population-based topic controversy" model in a sense that the controversy is observed from a given population and always requires some population, motivated by Dori-Hacohen's definition of controversy (Dori-Hacohen, 2017). Topic controversy models have been mainly studied within the medium of Wikipedia, social media, and Web queries. In later sections, we will review how existing work has captured conflicts and contention to

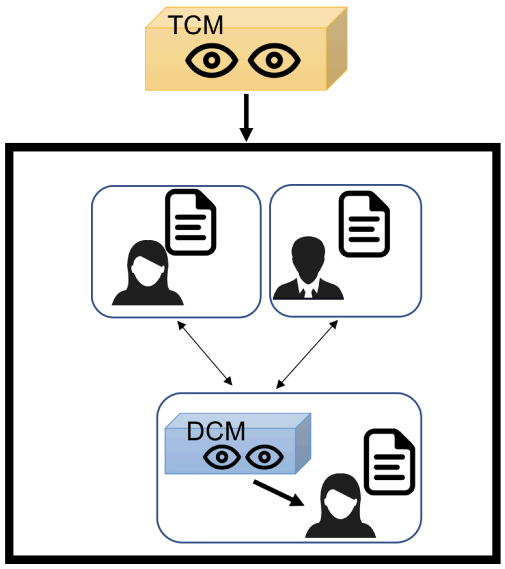

Figure 2.1: Topic Controversy Models (TCM) has a bird's eye view from outside of the discourse of the topic that can observe the interactions between people and their disputes. Document Controversy Model (DCM) has a view within a text entity (e.g., web documents, tweets) without being able to observe the interactions with the other entities.

identify controversy in each medium. We will categorize existing work by what type of signals they have captured.

2.1.2 Document Controversy Models

Document controversy models refer to controversy models that determine the probability that a given document object is controversial. Document controversy models differs in their nature from topic controversy models for mainly two reasons. First, while topic controversy models considers one topic, a document usually contains a mixture of topics. Specifics such as *which topics* are discussed and *how* it is discussed changes the probability that the document is controversial. Second, while a topic controversy model examines whether there has been a dispute in a given population, hence their observation is made in a meta-level of the entities (e.g., documents, people) and their interactions. In a document controversy model, a document object is one

entity that comprises a population. The key difference between the two models is where their "eyes" for observation locates: that topic controversy models take a bird's eye view on the discourse on the controversial topic, whereas document controversy models takes a perspective from one entity that participates in the online discourse. This is demonstrated in the Figure 2.1.

The document controversy models have been mainly studied within two medium: general Webpages and news articles. While most existing approaches are topic controversy models, document controversy models have been less studied, especially from a modeling perspective.

2.2 A Survey of Controversy Detection Algorithms

2.2.1 Detecting Controversy in Wikipedia

Wikipedia probably has been the most-studied medium for controversy because it has the advantage of having the entire edit-history available, which is user interaction log of how discourse of the topic has been developed. Kittur et al's work (2007) was a pioneering effort to characterize conflicts in Wikipedia and introduced the task of identifying articles with high conflicts. They demonstrate that the cost of coordination and conflicts is increasing at a global level in Wikipedia, meaning that while direct work on articles is decreasing, indirect work such as discussion and maintenance activity is increasing, which brings people's attention to understand and analyze these conflicts.

To identify articles with high conflict, they trained a SVM regression algorithm. As a subset of Wikipedia articles are manually labeled with a "*controversial*" tag by editors, they developed a metric called Controversial Revision Count (CRC), which is the number of "*controversial*" tags in the revision history of that article. Their regression algorithm was trained to predict CRC, which they treat as a proxy of the level of controversy of the given article. The features they used include the number

of reverts, the number of edits, the number of anonymous edits, which are intended to relate the level of conflict to the number of reverts between the two editors.

While Kittur et al.'s model was a supervised approach that requires manually-labeled data, Vuong et al. (2008) proposed a way to make the model unsupervised: instead of analyzing the actual article content, they modeled disputes from the interaction between two editors. They define a *dispute* between two editors as the number of words that have been deleted from each other in the article's edit history. In their model, an article is more controversial if it has more disputes between two contributors who are known to engage with less controversy. The authors also discover that some of the "disputes" were dedicated to eliminate vandalism. To address this issue, Yasseri et al. (2012) focus on distinguishing such vandalism from meaningful controversy, introducing *M score*, which we build upon for our work.

In Yasseri et al.'s work, they define a dispute as a "mutual revert" between two editors where the two revert each other mutually. As determining whether each dispute is a meaningful dispute or vandalism is crucial for correctly measuring the level of conflicts, they estimate the reputation of the two reviewers who participate in a mutual revert. The idea is to give more weight to the dispute between the two reviewers who are deemed to be trustworthy, while penalizing the one involving at least one reviewer who is less credible. Therefore, an article is more controversial when there are more mutual reverts between the two editors, in which both of them have higher reputation. A reputation of an editor is measured by the total number of edits that the editor has contributed to a given article.

Brandes et al. (2009) and Sepehri Rad et al. (2012) turn to a network structure to characterize conflicts in Wikipedia and analyze polarization of the community in the editor network. The intuition behind this is that a more controversial topic will likely have a more polarized editor network. They build a collaboration network where nodes correspond to editors and signed edges correspond to their positive or

negative interactions. Negative interactions can be defined in a few ways such as the number of deletes between two editors (Brandes et al., 2009; Sepehri Rad et al., 2012), the number of mutual reverts and the presence of negative terms in comments (Sepehri Rad et al., 2012). However, as the variance of the polarization score between controversial and featured articles that are popular and of high quality article is known to be high, its applicability is known to be limited. This is due to the fact that the positive interactions were not taken into consideration between the two editors while both negative and positive edges are known to be important in a signed network (Rad and Barbosa, 2012).

Sepehri Rad and Barbosa (2012) argue that a powerful controversy detection algorithm should have a high discriminative power and satisfy monotonicity. They performed a comparative study on the five existing controversy algorithms that utilize different features. In their evaluation, while a mutual-reverts based classifier (Yasseri et al., 2012) (M score) has less discriminative power than a meta-data based classifier (Kittur et al., 2007), it is the only classifier that satisfies the monotonicity criteria. Their monotonicity criteria defines that a controversy function should have less or an equal score to a given article if some parts of the article were removed from it. The authors explain that the intuition behind monotonicity is that removing some parts will only likely remove some of the disputes, hence it cannot increase the controversy level of that article. However, note that this is based on the assumption that the level of controversy is proportional to the number of disputes. One could argue that as more non-controversial content exists in the document, the level of controversy goes down. We will revisit later the fact that M score is not only monotonic within a given article but also over time because as the longer the edit history gets, the amount of mutual reverts get accumulated. We discuss that this does not accurately reflect the reality that controversy changes over time and propose a time-adaptable controversy score that changes over time rather than being cumulative.

Finally, we summarize the dispute signals and features used for controversy detection in Wikipedia in Table 2.1. We categorize existing work by how and whether it utilizes the four types of signals – disputes, meta-data of articles, the language of articles (e.g., keywords, n-grams of the article content) and a network structure of editors. – that are used to identify controversy in Wikipedia.

2.2.2 Detecting Controversy in Social Media

In an era in which new controversies rapidly emerge and evolve on social media, there have been numerous efforts that aim to analyze, characterize, and identify controversial topics from social media, particularly in Twitter. Popescu and Pennacchiotti (2010) were the first to pose the problem of identifying controversial events from Twitter and explore an extensive set of features such as linguistic and structural features, sentiments, and controversy features. Their controversy features include the ratio of mixed sentiments, the fraction of terms that are in a controversy lexicon, or controversy-indicative hashtags.

Conover et al. (2011) discover that the retweet network exhibits highly segregated communities for controversial topics. This important finding has motivated other subsequent work (Guerra et al., 2013; Garimella et al., 2016; Fraisier et al., 2017) to focus on the retweet network structure and model the polarization of the network as a key feature in the models for controversy detection on social media. Garimella et al. (2016) develop this model further to quantify how controversial the topic is by proposing a random-walk based measure between two partitioned-graphs (i.e., communities) from the retweet network. For the focus of their study, Garimella et al. make a simplifying assumption that there are always only two conflicting communities and that those two communities are of the same size, and uses a graph partitioning algorithm, METIS (Karypis and Kumar, 1998), which aims to cut the graph into two subgraphs of the similar size.

Colleto et al. (2017) focused on capturing local patterns of user interactions to identify controversial tweets by analyzing the reply threads. They construct two types of edges in an user graph, "reply" and "retweet", and use the patterns of dyadic or triadic relations as features for controversy classification. They discover that a pattern of two users where they do not follow each other but one replies to the other is the most useful feature that distinguishes controversy from non-controversy, whereas replies to someone he/she follows is not a relevant feature.

Fraisier et al. (2017) experimented with various community detection algorithms to identify user stances on Twitter. For the two topics of "Scottish Independence Referendum" and "US Midterm Elections", they attempted to predict user stances between two conflicting stances, such as "Favor vs Against" or "Democrat vs Republican". They discovered that the retweet networks are a generally better way to detect like-minded communities than mention graphs. On the retweet networks, algorithms that rely on information diffusion such as label propagation (Raghavan et al., 2007) and INFOMAP (Rosvall and Axelsson, 2009) were shown to be the leading ones. Based on the fact that INFOMAP finds communities based on the flow of information present in the network, they argue that in some way, stances "follow" the information on Twitter.

In our earlier work (which is not included in this thesis), we proposed a theoretical model to formally define controversy and argued that controversy is not a static universal value and is better measured with respect to a given population (Jang et al., 2017). Our model suggests that "contention" among people and "importance" of the topic to the people are the primary dimensions that contribute to the level of controversy. To compute contention, our model considers the size ratio of two groups of people who take each conflicting side on the controversial topic. We validate our model by analyzing a few controversial topics in social media. As a hashtag-based approach was studied as a high-precision method for collecting stanced-tweets by using

Table 2.1: Controversy detection algorithm in **Wikipedia** and features used. ✓ indicates that the corresponding feature is used and △ indicates that the feature was *indirectly* used.

Work	**Dispute signals**	**Meta-data**	**Language**	**Network**
Kittur et al. (2007)	-	✓	-	-
Vuong et al. (2008)	deletes	-	-	-
Brandes et al. (2009)	-	-	-	✓
Sepehri Rad et al. (2012)	mutual reverts, deletes, negative terms in comments	-	-	✓
Yasseri et al. (2012)	mutual reverts	-	-	-
Dori-Hacohen et al. (2016)	-	✓	△	-
Zielinski et al. (2018)	sentiments	-	-	-

a manually-curated hashtags (Mohammad et al., 2016c), we also manually curated stance-indicative hashtags (e.g., `#MAGA` to support Donald Trump, `#ImWithHer` to support Hillary Clinton in the 2016 US Presidential Election) for each topic and estimated the size of the communities of conflict from the tweets that use such hashtags. Our results demonstrate that they align well with reality by showing a spike in the level of controversy where we can easily find an external event that can explain this phenomenon. This hashtag-based approach further motivated our work in controversy summarization in social media in Chapter 7.

While dispute signals are the most prominent features that most existing work have utilized in Wikiepdia, a network structure that globally characterizes the segregation between the communities or locally characterizes the disputes between users has been understood as the most useful property to identify controversy in social media. We categorize the existing work by the three types of main signals, sentiment, language, and the network structure, that have been used to understand controversy in social media in Table **??**.

2.2.3 Detecting Controversy in Online News and Webpages

Identifying controversy in online news and webpages requires different models from the ones used to identify controversy in Wikipedia or social media, because they

Table 2.2: Controversy detection algorithm in **Social media** and features used. ✓ indicates that the corresponding feature is used.

Work	**Dispute signals**	**Sentiments**	**Language**	**Network**
Popescu and Pennacchiotti (2010)	-	✓	✓	
Conover et al. (2011)	-	-	-	✓
Guerra et al. (2013)	-	-	-	✓
Garimella et al. (2016)	-	-	-	✓
Jang et al. (2017)	-		✓	
Coletto et al. (2017)	-			✓
Fraisier et al. (2017)	-	-	-	✓

Table 2.3: Controversy detection algorithm in **Web pages and news articles** and features used. ✓ indicates that the corresponding feature is used.

Work	**Medium**	**Dispute Signals**	**Sentiment**	**Language**
Choi et al. (2010)	News	-	✓	
Mejova et al. (2014)	News	-	✓	
Dori-Hacohen and Allan (2015)	Webpages	-	✓	✓
Jang and Allan (2016)	Webpages	-	-	✓
Jang et al. (2016)	Webpages	-	-	✓
Beelen et al. (2017)	News	news comments		✓

usually do not have any structured meta-data or user interaction signal to identify controversy from, except for some work that considered the user comment thread in online news data. While the presence of polarization of a user interaction network or dispute signals have been studied to be useful signals to identify controversy from Wikpedia and social media, we have to rely on text analysis of the documents without extra features. Naturally, sentiment analysis of text is considered to estimate the features.

Choi et al.'s work (2010) was one of the pioneering work that investigates identifying controversial issues and subtopics from news articles using various features, particularly a mixture model of topic and sentiment. They define controversial issues as concept that invokes conflicting sentiment or views and a subtopic as a noun phrase that provides a reason that the issue has conflicting sentiment. They measure the level of controversy of a given phrase based on the topic importance and the difference

of the sentiment of the terms in it. They performed a qualitative analysis for their results.

While some past work uses sentiment as a signal when researching controversy, others have argued that opinion and controversy are distinct and non-overlapping concepts. Awadallah et al. (2012) explain that political controversies are much more complex and opinions are often expressed in subtle forms, which makes determining polarities much more difficult than in product reviews, in which sentiment analysis and opinion mining techniques have been used. Mejova et al (2014). argue that controversy and sentiment are not directly related.

Dori-Hacohen and Allan's work (2015) was the first attempt to extend the controversy detection problem to general webpages in an open-domain. They first investigate the usefulness of sentiment in identifying controversy in Webpages. They demonstrate that sentiment alone cannot be a good signal to identify controversy by showing that a sentiment analysis baseline fails to identify controversial topics in Wikipedia, which supports the claim from other work (Awadallah et al., 2012; Mejova et al., 2014).

They begin by generating a query from a web page, and retrieving the K nearest neighbors from Wikipedia. They create a binary classifier by aggregating controversy features that are computed in retrieved Wikipedia pages (Yasseri et al., 2012; Das et al., 2013),

There also have been a few attempts to detect controversial content with lexicons. Roitman et al. (2014) focused on a claim-oriented document retrieval task. They retrieve Wikipedia articles that contain relevant claims about a controversial query topic using manually-curated controversy lexicon. Mejova et al. (2014) use crowdsourcing to label controversial words.

Beelen et al. (2017) also studied identifying controversy from news articles by investigating extensive features that indicate controversy from the document text as

well as people's comments. They showed that their comment-based method that considers the meta-data of comments of the news articles, was more effective than a content-based approach that considers the text of the news articles for controversy detection in news articles.

2.2.4 Detecting Controversy in Search Queries

There has been little work done in finding controversial topics from search queries except for the work of Gyllstrom et al. (2011). They observed popular claims in search query log to identify controversial topics. Specifically, they create a *claim* search query that has a pattern of 'X [is/was/are/were] Y' to obtain an insight whether popular claim queries from a search engine contain conflicting sentiments. They send 'X [is/was/are/were]' to a search engine to obtain the top suggestions to find the claims. Among the claims, they observe whether a claim that is a negation of another claim exists, such as 'X is *fake*' and 'X is *real*'. When there is a pair of claim queries that negate each other, they determine that the entity in the claim is controversial. However, this approach is limited in several ways. First, they require abundant search query log for the approach to be effective. Second, their approach is limited to controversies that can be summarized in the form of simple claims using an adjective or a noun. There are many controversies that are too complex to be described as simple claims, and not all controversial claims necessarily have the negating claims. For example, a controversial claim has been raised whether Apple has purposely slowed down the performance of the old iPhones to accommodate their aging batteries (Nusca, 2017). Not only is this controversy too complicated to be written as a simple 'X is Y' type of claim but also negating claim was not raised from users.

2.2.5 Summary

Controversy detection methods have been studied within a given medium, mainly among Wikipedia, social media, webpages and news article. While how controversy, as

a complicated and subjective social phenomenon, should be computationally defined and characterized itself is still an open question, we summarize previous work by the type of features they captured to estimate controversy in each medium in Table 2.1, 2.2, and 2.3.

In Wikipedia, capturing disputes and conflicts between the editors has been the main focus of previous work. While how editor network is structured and the meta-data features of Wikipages have been also studied, the disputes between the editors have shown to be the most prominent signals characterize controversy in Wikipedia.

However, existing approaches focus on analyzing "present" dispute signals on Wikipedia, which leads them to be a precision-oriented approach than a recall. For example, topics that are less popular tend to get less edits in general, hence seemingly less controversial than they actually are. We address this issue in Chapter 4 and Chapter 6 to improve Wikipedia-based controversy approaches to be more reliable.

In social media, controversy has been mostly characterized by how strongly people with similar opinions form a community on a controversial topic rather unlike explicit conflicts in Wikipedia. In Wikipedia and social media, the existing approaches proposed fall into the category of topic controversy models.

On the other hand, Web pages and news articles differ from the other mediums because they do not have auxiliary information such as conflict history between users user interaction behaviors, and the focus of their problem is to judge the controversy of a given object, they use document controversy models. The main signals that have been studied are sentiment and the text of the document to find topics. One type of model is used within another model. The state-of-the-art algorithm (Dori-Hacohen and Allan, 2015) uses Wikipedia controversy models to identify controversy in the document bu using similar Wikipedia topics from the document.

Existing sentiment-based algorithms to find controversy in documents are mostly lexicon-based approaches where they look for matching keywords from the predefined

lexicon list. Such approaches are not scalable and limited as in sentiment does not always reflect controversy. Therefore, we investigate a probabilistic approach to use the language of the document to estimate the probability later in Chapter 5.

2.3 Detecting Subjectivity and Bias

Cartright et al. (2009) attempted to characterize subjectivity in Web documents by proposing two new metrics, *provocativeness* and *balance*, which could suggest the document's topic is controversial. They define the provocativeness as the average level of subjectivity of all relevant units (e.g., documents) to the topic and the balance as the amount of imbalance between the negative and positive opinions of a topic. They applied the two metrics to characterize the topics from TREC Blog Track and presented an analysis that the topics used in the blog track tend to be provocative.

As controversial topics are likely to use biased language with regard to a certain stance that the author takes on the given controversial topic, bias is often related to controversial topics. To identify biased language from text, Recasens et al. (2013) discovers two classes of biases, *framing bias*, which injects a certain perspective and subjectivity, and *epidemiological bias*, which is related to truthfulness of the statement. Between the two, framing bias is more closely related to the controversial topics. They observe that framing bias occurs when subjective intensifiers or one-sided terms are used, which reveal the author's stance on the given topic.

While one-sided terms are more closely related to controversy, such terms are topic-dependent and difficult to obtain as it require stance detection on corpus. Previous literature has focused on a two-way classification of classifying the author's stance to two conflicting stances such as "support" vs. "against" or support "Donald Trump" vs "Hillary Clinton"). For stance classification, the subjectivity of language and sentiment lexicons were considered as features as well as unigrams, bigrams, distributional similarity, etc (Recasens et al., 2013). In Ricasens et al.'s work, Riloff et

al.'s work where the linguistic patterns that indicate a subjectivity in a sentence were used as part of the features to capture bias. We will discuss about research effort in stance detection focusing on social media in Section 7.2.

2.4 Explaining Controversy on Social Media

Explaining controversy is a relatively new area and there has been little prior work on this problem. In our work, we focus on explaining the two conflicting stances that make up controversy. For this problem, two research areas are mainly related, stance detection and summarization on social media.

2.4.1 Stance Detection on Twitter

In order to find tweets that represents each conflicting stance for a summary, stances identified in each tweet would be an useful knowledge.

Stance classification on Twitter usually consists of two main tasks: (1) classifying the text's stance (against, favor, or neutral) given a topic, and (2) classifying the twitter users' stances. The former task drew attention when 2016-SemEval Task 6 released a dataset of tweets with stance annotations (Mohammad et al., 2016b). The results of various approaches were shared after the competition (Mohammad et al., 2016c), and later more successful approaches were proposed including one that uses a bi-directional conditional LSTM for classifying the stance and opinion target on Twitter (Augenstein et al., 2016). For the latter type of task, Johnson and Goldwasser developed a method to classify stances of politicians on Twitter using relational representation (Johnson and Goldwasser, 2016).

The 2017 Fake News Challenge Stage 1 (FNC-1) shared task focused on a stance detection task as a crucial first step towards fake news detection (Pomerleau and Rao, 2017). In this task, an input is given as a headline and a body text either from the same news article or two different articles. Then an algorithm should classify the

stance of the body text with respect to the claim made in the headline into one of four categories – "Agrees" (the body text agrees with the headline), "Disagrees" (the body text disagrees with the headline), "Discusses" (the body text discusses the topic of the headline but does not take any stance), and "Unrelated" (the body text discusses a different topic from the headline).

Because the stance detection in this task deals with a longer document than a tweet, it poses a new challenge from the stance detection in tweets. In tweets, the challenge comes from the fact that short text give little hint and context for identifying a stance. On the other hand, a long document may contain statements that suggest one stance when considered in isolation, but imply the opposite stance given the context of the document. The top ranked FNC system was from Talos Intelligence's SLOAT in the SWEN team, who used a weighted average model of a deep convolutional neural network and a gradient-boosted decision tree model. For their decision tree model, they used word count, TFIDF, sentiment, and singular-value decomposition features with the pre-trained word2vec embedding (tal, 2017; Hanselowski et al., 2018).

From these recent two stance detection shared task, one common lesson we learned is that the investigated stance detection task is a difficult problem. In the SemEval 2016 Stance Detection in Tweet share task, none of the participating team consistently outperformed the baseline. Hanselowski et al. 2018 analyzed the top-performing approaches in FNC-1 share task and concluded that more sophisticated machine learning techniques that have a deeper semantic understanding are needed as the best performing features are not yet able to resolve the difficult cases yet. However, we argue that while stance detection is closely related to our problem, our goal is not to accurately classify the stances of all tweets. Our problem is also more robust to misclassification errors of stances as we can take the tweets with highest stance confidence as part of the summary.

2.4.2 Twitter Summarization

There has been much work on summarizing Twitter postings through most of them focuses on summarizing events (Sharifi et al., 2010; Duan et al., 2012; Chakrabarti and Punera, 2011; Inouye and Kalita, 2011; Yulianti et al., 2016). Inouye et al. 2011 compare multiple summarization algorithms for Tweet data, and their extensive experiments suggest that the SumBasic algorithm (Nenkova and Vanderwende, 2005) produced the best F1-result in human evaluation. SumBasic is a summarization algorithm that uses the term frequency exclusively to create summaries. As a simple system based on word frequency in the document set, SumBasic outperformed any other complex system at the time. SumBasic computes the best k posts from the input documents that contain a lot of high frequency terms. We choose SumBasic as our baseline method.

Some work has focused on generating contrastive summaries from opinionated text (Paul et al., 2010; Guo et al., 2015). Particularly, Guo et al. studied tweet data to find a controversy summary. They find a pair of contrastive opinions by integrating manually-curated expert opinions and clustering the pairs to generate a summary. However, their model needs curated expert opinions, which requires constant human effort to maintain as the topic evolves.

CHAPTER 3

PROBABILISTIC MODELING OF CONTROVERSY DETECTION

3.1 Introduction

This chapter discusses a probabilistic framework for the task of detecting controversy of a given web document. Dori-Hacohen and Allan (2013) first introduced the problem of detecting controversial topics in Web documents. The goal of this task is to make a binary classification on whether or not a given document discusses controversial topics. Dori-Hacohen and Allan proposed a k-nearest-neighbor (kNN) classification approach for this task and conducted a proof-of-concept experiment. Their pilot study demonstrates that given a query document, identifying similar k Wikipedia pages and their controversy levels can effectively identify controversy in the document. They first mapped each query webpage to k related Wikipedia pages (Wikipages) that are manually identified, and used the annotated controversy level of the selected Wikipages to produce a final controversy score for the document. Later, they proposed the first fully-automated algorithm that implements the kNN approach (2015), which we call "kNN-WC algorithm".

The kNN-WC algorithm has been shown to be effective. However, its lack of theoretical underpinning leaves a gap between our theoretical and empirical understanding in this problem. While the kNN-WC algorithm is an implementation of the underlying kNN approach, the algorithm adopts a few assumptions that were not specified in the kNN model. Although Dori-Hacohen and Allan leave the theoretical groundwork of the kNN-WC algorithm largely unexamined, we propose that the algorithm has been implicitly instantiated from an underlying probabilistic model, which

we name as kNN-WC model. We aim to derive the probabilistic kNN-WC model that can explain the assumptions and the behaviors of this algorithm in a more general sense.

Why do we need a model when we already have an algorithm that works reasonably well? When an algorithm is instantiated from a theoretically-grounded model, we can obtain a better intuition of why the algorithm works. Having a model allows us to understand mathematical foundations and to evaluate a set of assumptions made to design the algorithm. This can help us to challenge the existing assumptions and develop better algorithms.

We therefore analyze and derive a model for the kNN-WC algorithm. Our goal here is *not* to design a new model but instead to derive a probabilistic model that explains the kNN-WC algorithm. Deriving a probabilistic model for the state-of-the-art algorithm sets a path for us to investigate controversy detection task in a probabilistic framework. We identify the assumptions that the kNN-WC algorithm made beyond the underlying KNN approach and resultant properties that the model has. We later demonstrate that deriving such a model can be used to extend the approach and design models and algorithms with substantially improved efficiency, accuracy, and generalizability. Specifically, deriving this model is a crucial step towards understanding this problem in many ways because it allows us to answer the following research questions:

- **Theoretical Understanding of the Problem (Chapter 3):** What is the mathematical background of the model and what assumptions were made in the model?

- **Revisiting the algorithm (Chapter 4)**: How reasonably did the algorithm implement the assumptions of the model? How accurately do the heuristics adopted by the algorithm estimate certain probabilities? Are there better ways to estimate them?

- **Testing a new hypothesis (Chapter 5):** What are the drawbacks of the assumptions and what could be an alternative model that has different properties than the given model?

We know of only two efforts to examine a theoretic model for controversy (Dori-Hacohen, 2017; Zielinski et al., 2018). Because both of the proposed models computationally define controversy as the disputes within a given community (or a 'population'), they require auxiliary signals of disputes to estimate controversy, such as Wikipedia's edit history or user interaction behaviors on social media. Therefore, those models are not directly applicable to Webpages that do not have any external signal but just text.

3.2 Background: Theoretical Models to Define Controversy

While there has been little work toward developing theoretical models in the domain of controversy, we introduce two related efforts that have modeled controversy. Dori-Hacohen (2017) presented a theoretical model to define controversy within a group of people, or a *population*. Her model is inspired by growing disparity between scientific understanding and public opinion on certain controversial topics, such as climate change, evolution, and vaccination. While many scientists think that there is no controversy with regard to those topics, in a general population, non-scientific claims and arguments proliferate causing the topics to be highly controversial. Hence, she argues that controversy is not a global and static value for a topic, but rather defined by a function that takes a given population as well as the topic.

Let Ω be a population of n people. Let T be a topic of interest to at least one person in Ω. Her model assumes that controversy is a multi-dimensional factor of traits that can be observed in Ω. She hypothesizes that such dimensions include *contention* to measure how contentious the topic is, *importance* to measure how important the topic

is to people, and *conviction* to encode who strongly holds their belief in their stances as follows:

$$controversy(\Omega, T) = f(contention(\Omega, T), conviction(\Omega, T), importance(\Omega, T))$$

Dori-Hacohen defines the probability of contention within a population as the probability of randomly drawing two people that have different stances that are in conflict with each other on a given topic. While she modeled "contention" in her work, she left other dimensions unexplored. In work outside of this proposal, we explored the dimension of "importance" by suggesting that the importance of the topic should also be measured with regard to the population, specifically by the ratio of people who are affected by T in Ω (Jang et al., 2017). This was measured by counting people who post tweets on the topic at least once during the time of observation.

Zielinski et al. (2018) later also recognized the necessity of having a conceptual model that formally defines controversy, which supports our definition of controversy. Their work is based on a Merriam-Webster dictionary definition of controversy as an "*argument that involves many people who strongly disagree about something: strong disagreement about something among a large group of people.*" Their proposed function takes three variables, a given object d (e.g., a webpage, a Wikipage, search queries), a given community Ω, and an empirical distribution of opinions given by members in Ω in d (E^d_Ω), to output a binary classification as follows:

$$f(d, \Omega, E^d_\Omega) = \{\text{controversial, non-controversial}\} \tag{3.1}$$

Although they used a slightly different terminology such as referring to *population* in Dori-Hacohen's model as *community*, the underlying assumption of the model captures the same intuition that "contention" within a given set of people is the main feature to measure controversy of a given object or topic itself. In this thesis, we will

use the term "population". While this model shares the same goal with our model, they assume that there is a community attached to the query object where disputes can be observed from.

3.3 A Probabilistic Model of the knn-WC Algorithm

In this section, we analyze and derive a model for the kNN-WC algorithm. We stress that our goal is not to design a new model, but to propose a theoretical model that explains the kNN-WC algorithm. The kNN approach proposed by Dori-Hacohen and Allan (2013) for controversy detection takes the following steps:

1. **Finding k similar topics:** When a webpage is given as an input, it finds k nearest-neighbor similar topics.

2. **Identifying the level of controversy for the k topics:** For the k similar topics, it identifies the level of controversy of each topic.

3. **Classify:** Based on the level of controversy on the k topics, it aggregates them to finally classify whether or not the query document is controversial.

As this approach has been shown effective, they proposed a fully-automated implementation of the kNN model, named the kNN-WC algorithm (2015). We summarize the kNN-WC algorithm as the following four steps:

1. **Retrieving k Wikipages via a document query:** When a webpage is given as an input, they find k nearest-neighbor Wikipages by generating a query of keywords extracted from the document.

2. **Computing controversy score on Wikipages:** From each of the k Wikipages, they automatically computed three controversy scores: C score (Das et al., 2013), M score (Yasseri et al., 2012), P score (Dori-Hacohen, 2017). In addition to these, they extracted D score that is a binary score that indicates the presence of **D**ispute tags assigned by Wikipedia editors (Kittur et al., 2007).

3. **Aggregate:** They aggregated the multiple scores of k Wikipages using average or max operators.

4. **Vote and classify:** They apply a voting scheme to turn the aggregated scores into a final binary decision, controversial or non-controversial.

Let us define a probabilistic framework that explains those steps by estimating the probablistic components. Let D be the text of document, and T be the topic of the document D. In this model, a topic is defined by a Wikipedia page (Wikipage) including its meta-data such as edit history. For example, T would be the most relevant Wikipage to D from the set $\mathcal{W}$ that contains all possible topics (i.e., Wikipages). We will interchangeably use the term topics and Wikipages from this point.

Finally, we define C be the binary variable to denote the controversiality of D. $P(C = 1|D)$ indicates that D is controversial, and $P(C = 0|D)$ means the opposite. For simplicity, we define the constant variable c to denote $C = 1$ and represent the query probability in a concise form: $P(c|D)$ to denote the probability that D contains controversiality and $P(nc|D)$ to mean that D does not contain controversiality (i.e., contains non-controversiality). The model aims to estimate $P(c|D)$ to determine whether or not the given document D contains controversiality. We summarize the notations used in our modeling in Table 3.1.

Table 3.1: A summary of notations used in our probabilistic framework

Symbol	Meaning
D	A document text consisting of words
T	A topic of D. In this model, a Wikipage.
$\mathcal{W}$	A set of all topics. In this model, Wikipedia pages
C	A binary variable to denote a D contains a controversiality
c	A constant to denote that $C = 1$
$P(c\|D)$	$P(C = 1\|D)$, the probability D contains controversiality.
$P(nc\|D)$	$P(C = 0\|D)$, the probability D does *not* contain controversiality.
Ω_w	A set of Wikipedia editors who contributed Wikipage w
q_D	A query generated from D

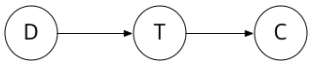

Figure 3.1: A simple Bayesian network kNN-WC model is based on. D: Document, T: Topic, and C: Controversiality.

First, we interpret "D containing controversiality" to mean that D discusses a *controversial topic*. $P(c|D)$ can be obtained from a marginal probability of the joint probability $P(c, D, T)$ for all possible topics of w in $\mathcal{W}$.

$$P(c|D) = \frac{P(c, D)}{P(D)} = \frac{\sum_{w \in \mathcal{W}} P(c, D, T = w)}{P(D)} \tag{3.2}$$

Because the probabilities $P(c|D), P(T|D)$, and $P(c|T)$ are closely associated with each other, we represent their relationship with a probabilistic graphical model that has three random variables, D, T, and C. We capture the following algorithm's flow by constructing a linearly-structured Bayesian network as shown in Figure 3.1: the topics (T) are determined by the query from the document (D), the controversiality (C) is determined by the contention level of topics. Intuitively, if the topic of the document is known, controversiality can be derived from that topic, which explains why C and D are conditionally independent given T. Based on the network, a joint probability distribution, $P(c, D, T)$ is defined as follows:

$$P(c, D, T) = P(c|T) \cdot P(T|D) \cdot P(D) \tag{3.3}$$

Finally, we derive $P(c|D)$ from Eq. (3.2) and Eq. (3.3). $P(c|D)$ is broken down to two components, the probability that a given document D retrieves a topic T, and the probability that T is controversial. For estimating $P(w|D)$, instead of considering all of D, they generate a query q_D from D to retrieve w. In addition, instead of considering all Wikipedia pages to aggregate the probabilities from, they take the top k most relevant Wikipages and estimate the probabilities from them. Let the top k most relevant Wikipages of D as $\mathcal{W}_D$:

$$P(c|D) = \sum_{w \in \mathcal{W}_D} \big[P(c|w) \cdot P(w|q_D)\big] \tag{3.4}$$

3.3.1 Estimating $P(c|w)$ using Contention

In our model, $P(c|w)$ indicates the probability that a given Wikipage w is controversial. There has been some work that focused on estimating the level of controversy of Wikipages. For the kNN-WC algorithm, three state-of-the-art techniques (Yasseri et al., 2012; Dori-Hacohen, 2017; Das et al., 2013) as well as binary dispute tags that are manually-curated by Wikipedia editors have been tested. We call them Wikipedia Controversy Features (WCF). Among these, M score has been shown to be most effective for their framework. Therefore, we discuss M score as well as P score that captures the same intuition as M score but that is derived from a more probabilistic grounds.

P score (Dori-Hacohen, 2017) and M score (Yasseri et al., 2012) both measure controversy as the level of "contention" within a group of people. This viewpoint is proposed to define controversy by the population model. While P score is an application of the population model in Wikipedia, the viewpoint retrospectively[1] explains the intuition behind M score.

Recall that the population model argues that the level of controversy of a given topic can only be answered with respect to a given population, and specifically, with regards to how *contentious* the topic is within the *population*. Both P score and M score assume that a population was given as the set of Wikipedia editors who contributed to the given topic. We explicitly transform the query $P(c|T)$ to an equivalent population-aware query by treating Ω_w, a population of Wikipedia editors on Wikipage w as a given parameter when w corresponds to the topic T.

[1]The controversy-population model was proposed 5 years later than the M score.

$$P(c|T) = P(c|w) = P(Contention|w; \Omega_w) \tag{3.5}$$

To estimate the level of contention, M score and P score both use "mutual reverts", online activities of Wikipedia editors where two editors have reverted each other's contribution, as a sign of disputes. The common intuition that both measures try to capture is that the contention increases as there are more reliable mutual reverts.

We first denote a set of Wikipedia editors that have contributed to a Wikipage w as $\Omega_w = \{p_1, p_2, ...p_n\}$. We define $mutualrevert(p_i, p_j)$ as a binary relationship that indicates whether reviewers p_i and p_j have mutually reverted each other. However, not all mutual reverts are meaningful. Vandalism is an act of maliciously editing Wikipages. Some mutual reverts are caused to fix these malicious activities, and should not be counted towards measuring contention.

Let $MR_D = \{(p_i, p_j)|p_i, p_j \in \Omega_w, \text{s.t.}, i < j \wedge mutualreverts(p_i, p_j)\}$ be the set of unique pairs of editors that have mutually reverted each other on D. Sumi et al. (2011) define $N_{p,D}$ be a reputation score of editor p, which indicates how credible p is (we omit the details here). The higher the reputation score is, the less likely p is to be a vandal.

M Score: To estimate if a given mutual revert is not caused by vandalism, they use a heuristic, $min(N_{p_i,D}, N_{p_j,D})$, to indicate how unlikely it is that any of the editors are vandals. M Score is computed as follows:

$$M = |\Omega_w^R| \cdot \sum_{(p_i,p_j)\in MR_D} min(N_{p_i,D}, N_{p_j,D}) \tag{3.6}$$

where Ω_w^R is a sub-population of Ω_w that is involved in at least one mutual revert that occurred in w. Since M score is not a probability, but an unbounded integer. We convert M score to a probability score by normalizing by the maximum M score among all Wikipedia pages.

P Score: Dori-Hacohen (2017) defines P score as the probability of drawing two random editors and the two editors have a mutual revert. Each mutual revert is discounted by the probability that each editor is not a vandal:

$$P = \frac{1}{|\Omega_w|^2} \cdot \sum_{(p_i,p_j)\in MR_D} \frac{N_{p_i,D}}{N^D_{max}+1} \cdot \frac{N_{p_j,D}}{N^D_{max}+1} \tag{3.7}$$

where N^D_{max} is the maximum reputation score of any editor who contributed to D.

By using the estimated probability from P or M score, we can develop the model as follows:

$$P(c|D) = \sum_{w\in\mathcal{W}_D} \big[P(contention|w;\Omega_w) \cdot P(w|q_D)\big] \tag{3.8}$$

Given this model, the kNN-WC algorithm makes a few approximation for the purpose of binary classification in a way that it uses cut-offs to turn the probability score into binary labels. First, based on the principle of a kNN classification, it considers the top k Wikipages instead of all pages. The kNN-WC algorithm chooses to aggregate the controversy scores of k topics via an average or a max aggregator. While the average aggregator more directly fits in our model, they show that the max aggregator is another heuristic that empirically works well. They also use a threshold for the controversy scores to turn them into binary flags for voting. While the kNN-WC algorithm makes effective choices for the purpose of the binary classification, the derived probabilistic model presents a discriminative power by being able to measure the level of controversy. It also suggests that alternatively, the level of relevance of the topic to the query document can be weighted differently to the final level of controversy as well as the level of the controversy level of each topic.

3.4 Discussion

From deriving the model of kNN-WC in Eq. 4.3, we learned the following properties about the kNN-WC model.

P1: The model has a population-based topic controversy model as a sub-component.

Because the level of controversy is determined from a topic-controversy model (Eq. 4.3), kNN-WC inherits the limitations that population-based controversy models generally have: it assumes that finding the evidence of "dispute" between people is a necessary condition for identifying controversy, while in reality, disputes are sparingly observed. For example, even for highly controversial topics, disputes are not observed constantly for all times as the human attention naturally is limited. The disputes are likely to be observed again when a new event spikes the interests. Topics that are controversial but less popular also suffer from the lack of dispute signals because it simply did not receive enough attention to generate contentious discussions for. Lastly, there are many similar topic instances (e.g., news articles, Wikipiedia pages) and it is impossible for all instances to show the same level of high disputes even for controversial topics. For example, we don't see the same level of disputes for all news articles on the comment section on the same controversial topic. We cannot simply expect all the articles to receive enough attention to generate a contentious discussion. Therefore, we need to ensure that errors caused from dispute sparsity are not propagated to the final prediction.

P2: Non-controversiality is not directly modeled.

kNN-WC is tuned to capture controversial signals by adding the controversy scores from each topic. When the document is non-controversial, the model expects to catch its non-controversiality because the topics retrieved would be non-controversial and contributes zero or small number of score values to the final score. However, the non-controversial topics only act in a way that it does not increase the probability that the document is controversial. It does not significantly differentiate the two cases where one mainly talks about the controversial topics and the other one mostly talks about non-controversial topics but briefly mention the controversial topic as a

passing comment. Theoretically, the latter case is penalized because the relevance to the topic should be lower when the given topic is not the main theme of the document. However, it is still susceptible to lean towards "controversial" because it is likely to contribute more to the final score than a non-controversial topic would with its high contention score. This is hinted from that in Eq. 4.3, the probability is a summation of non-zero components in the retrieved topic list. In fact, this issue is aggravated in the kNN-WC algorithm, a specific implementation that Dori-Hacohen and Allan proposed, when they treat the top K retrieved topics to have the same relevance probability. As long as the highly-controversial topic is retrieved in the top K list, the document's probability of being controversial is highly likely to be overrated.

kNN-WC adopts a principle that if there are controversial topics mentioned in the documents, it is likely to be controversial and the model is ready to "listen" to the controversial signals that is present in the documents, However, alternative principle could be considered: even if the controversial topics are mentioned, if the document mostly discusses non-controversial topics, the probability of controversilaity should be decreased. Perhaps, the balance of the controversial and non-controversial content could be considered.

P3: A documents' text is only a proxy to find topics.

In this model, the document's text is only considered as a proxy to find topics. The intuition of the model is that the controversiality of a document is determined by its latent topics. The graphical model behind kNN-WC model suggests that once the document's topic is given, the text of the document does not affect the probability that the document is controversial anymore via the conditional independence assumption. In other words, the documents' text is only used to identify the topics and it does not directly affect the probability that the document is controversial. In another model, alternatively, we could consider documents' text directly to estimate the probability of controversiality.

In the next two chapters, we will revisit the above three properties. In Chapter 4, we revisit the kNN-WC algorithm, a specific implementation of kNN-WC model. We show that the empirical performance of the algorithm is bounded by how realistically two probability components in Eq. 4.3 are estimated, and particularly limited by the issues presented in P1. We then propose methods to fix them to improve the algorithm. In Chapter 5, we propose Controversial Language Model (CLM), which addresses all P1, P2 and P3. We finally compare CLM and kNN-WC model in their empirical performance and via a qualitative analysis.

CHAPTER 4

REVISITING AND IMPROVING WIKIPEDIA-BASED CONTROVERSY DETECTION

We discussed in the previous chapter (Section 3.3) that the kNN-WC algorithm can be viewed as an instantiation of the probabilistic model presented in Eq. 4.3. In this context, while the kNN-WC *model* specifies the general probabilistic components, we use "kNN-WC *algorithm*" to refer to a specific implementation of the model, including how the probabilistic components are chosen to be estimated as proposed by Dori-Hacohen and Allan (2015). From the derived model, we discovered that any implementation of kNN-WC model should satisfy the following two assumptions:

> **A1:** $P(w|q_D)$ **assumes that a query generated from the document retrieves Wikipages that represent the document's topics.**
>
> **A2:** $P(contention|w;\Omega_w)$ **assumes that Wikipages that discuss controversial topics will show a high level of contention among the editors of the page, and vice versa.**

In this chapter, we revisit the kNN-WC algorithm and discuss how each assumptions often fail to be met in the current algorithm. We propose two solutions to improve the accuracy of each probability to implement the kNN-WC model more accurately.

4.1 Revisiting the assumptions for the kNN-WC algorithm

4.1.1 The Limitation of a Single Document Query

kNN-WC model assumes that a query generated from a document retrieves k relevant Wikipages to estimate the level of controversy from. To generate a query for the

document (i.e., a document query), the kNN-WC algorithm takes a straightforward solution of simply using the "best" k keywords. In the algorithm, they used the top k frequent terms.

However, we observe that generating a single global query from a document for retrieving relevant Wikipedia pages inherently brings two issues. First, as the document almost always contains multiple sub-topics, the generated query contains an unknown mixture of different sub-topics. This makes the query's intent less clear, as it targets many sub-topics at the same time and in unknown balance. Second, it is unlikely that all sub-topics are covered in the query – or covered appropriately – because keywords are extracted from a bag-of-words, which does not model the existence of sub-topics as it is. To address this issue, we investigate an alternate way of query generation, namely TILEQUERY: generating multiple local queries from topically-segmented documents (i.e., tiles) and aggregating multiple ranked lists from each query. We discuss this approach in Section 4.2.

4.1.2 The Limitation of Wikipedia Controversy Features

To estimate the level of controversy of a Wikipage, Dori-Hacohen and Allan examined previous work that studies the signals of dispute in Wikipedia (Kittur et al., 2007; Das et al., 2013; Yasseri et al., 2012). We refer to these signals as *Wikipedia Controversy Features (WCF)*. The algorithms that were used to generate WCF use meta-data of Wikipages, dispute signals in the page's edit history, or manual dispute tags assigned by Wikipedia editors.

The kNN-WC algorithm uses WCF to estimate $P(Contention|w, \Omega_w)$ because WCF is inspired by algorithms that model "edit-wars", the evidence of multiple editors (Ω_w) exchanging opposing opinions on the given Wikipage (w). We introduce the three features used in kNN-WC algorithm, which we also use for realization of our new model later:

C score This score was generated by a regression-based method (Kittur et al., 2007) that estimates the revision count of controversial Wikipedia pages, which are labeled with {controversial} tags. The algorithm was trained with edit-history information, such as the number of unique editors and number of reverts, as well as some metadata of Wikipedia pages . The score is normalized so that it ranges between 0-1.

M score Another controversy score studied by Yasseri et al. (2012) is based on statistical features of edits, which signify how fierce the "edit war" is. The statistical features include the number of mutual reverts of two editors, the number of editors participating in this edit-war, and the editor's reputation. M score is theoretically unbounded ranging from 0 to a few billions.

D score This is a Boolean value indicating whether a Wikipedia page contains a *dispute* tag in it. This tag is assigned by the page's contributors if the Wikipage's talk page shows some level of dispute. Unlike the above two scores, this label is manually curated. Hence, this score is extremely sparse; only 0.03% of the articles have a positive D score (Kittur et al., 2007).

Unfortunately, these approaches are all limited for the same reason: many Wikipages with controversial topics do not have sufficient edit-history to form an edit-war or the relevant edit-war has been delegated in other pages on the similar topic. There is a tendency that the heat of the edit-wars be focused on one Wikipage of a general and broad topic, leaving other related but sub-topical pages less attended. After all, there is simply no point of having the same "war" on all similar Wikipages. Table 4.1 shows an example of a few "abortion" related topics and their M and C score. While the "Abortion" page received a lot of attention, other pages with more specific topics such as *Abortion in certain countries* and *Abortion Act* had virtually no edit-wars. Unless there is a specific issue or event specifically tied to the page, all general disputes on

abortion have been delegated to the "Abortion" page. In other words, not having the "edit-war" does not necessarily mean that there was no war in this topic, but that the war has been happening somewhere else instead. This phenomenon causes the algorithm to easily make false negative errors (i.e., classifying "controversial" as "non-controversial").

Table 4.1: An example of M score and C score for Wikipages on "Abortion" that most sub-pages on "Abortion" have controversy scores close to 0.

	M score	**C score**
Abortion	4,102,593	0.300
Abortion_Act	0	0
Abortion_in_China	0	0
Abortion_in_England	0	0
Abortion_in_the_US	0	0.002

For the two limitations discussed, we propose two modifications in the framework, each of which tackles one issue.

4.2 Solution 1: Improving Document Topic Retrieval by Local Queries

The kNN-WC algorithm finds relevant Wikipages for a given query webpage by generating a query from the document. Querying By Document (QBD) is a well-motivated problem of finding other related documents for a given query document. There are numerous applications in real life where users can benefit from QBD: for example, research problems such as patent retrieval that returns similar patents to a new patent application, blog retrieval that finds related blog postings to a text document, and citation retrieval that finds related articles to an academic paper have all been studied (Kim and Croft, 2014; Yang et al., 2009; El-Arini and Guestrin, 2011).

Compared to traditional user queries, the main challenge of QBD stems from the fact that a document usually contains more and more aspects (i.e., sub-topics) as it becomes longer. If the document contains heterogeneous topics, the retrieved results should also contain heterogeneous topics. However, whether the query used to retrieve that list itself should be heterogeneous is questionable. We explore the interaction between a single query that models the entire document and a set of queries intended to capture each of the sub-topics of the document.

One straightforward solution for generating a document query is simply to use the "best" k keywords. However, generating the global keyword query from the document has two issues. First, as the document almost always contains multiple sub-topics, the generated query would contain an unknown mixture of different sub-topics. This would make the query's intent less clear, as it targets many sub-topics at the same time and in unknown balance. Second, it is unlikely that all sub-topics are covered in the query – or covered appropriately – because keywords are extracted from a bag-of-words, which does not model the existence of sub-topics as it is.

We consider a text-segmentation based query generation approach to address these issues. To generate a query of clear intent focusing on one sub-topic at one time and cover all present sub-topics, we model the document as a bag-of-*tiles*, where "tile" refers to a segment of text, similar to "TextTile" in the TextTiling technique (Hearst, 1997). In this model, we first segment the document into multiple tiles. Each tile is intended to contain fewer sub-topics than the document, ideally one sub-topic per tile. We generate a query from each tile and then aggregate the ranked lists obtained from the tiles. This can be viewed as a "divide-and-conquer" approach for document query generation.

Tiling the document for query generation is motivated by a general process of how documents are written. People tend to write a paragraph on a coherent sub-topic and have sub-topics flow in the document. Text segmentation is a relatively light-weight

way of considering sub-topics. Although topic modeling (Blei et al., 2001) can also be used to learn the sub-topics in a document, those topics are best learned from a corpus and are expensive to train as the collection grows. Hence, linearly segmenting the document is not only computationally efficient but also has the advantage of preserving the document structure property.

4.2.1 Related Work

Various approaches have been introduced to generate queries from a document as a whole. Smucker and Allan (2006) studied this find-similar items problem extensively. One of their valuable findings is that extracting a query from the document performs better for finding similar items than simply using the document alone as a query.

Keyword-based approaches assume that a good query from a document would be keywords that best summarize the document. A simple approach is to take k terms with the highest term frequency (TF) or TF·IDF score. Other popular term ranking functions include mutual information, KL divergence, and the χ^2 test. The RelevanceRank algorithm (Yang et al., 2009) constructed a Wikipedia graph with phrases extracted from the webpage and then identified keywords using a random surfer model.

Retrieval-based approaches use relevance feedback or pseudo-relevance feedback results to identify keywords. Queries can be iteratively refined by adding more terms from the top-ranked documents, and the newly modified query is issued again to obtain a new feedback list. The Rocchio formula and RM3 are used most popular for this task. In the patent retrieval domain, algorithms also use pseudo-relevance feedback (Ganguly et al., 2011b). Using an initial patent query, it obtains top-ranked documents and then formulates queries by selecting the sentences in the original document that have more likelihood given those pseudo-relevant documents.

Learning-based approaches use machine learning algorithms to learn keywords. Lee and Croft (2012) extracted important noun phrases and named entities and trained a CRF model given a user-specified passage in a document. This model uses various features such as Web n-gram, query logs, Wikipedia titles, and so on. However, their graph model does not scale well to a longer passage, such as a document. Kim et al. (Kim, 2014; Kim and Croft, 2014) used both pseudo-relevance feedback and machine learning technique. They trained a decision tree and used it to generate Boolean queries. From a baseline query extracted from a query document, it takes the top k pseudo-relevant documents and beyond k non-relevant documents as training examples and trains a decision tree to generate multiple Boolean queries. They then rank the queries to suggest top k queries to the user.

The closest existing work to using text segmentation for query generation is Ganguly et al.'s work on query reformulation (Ganguly et al., 2011a). They suggest that to reformulate a given query to increase its specification on the particular topic compared to the previous query, the terms from the document segment with the maximum number of matching terms can be added.

4.2.2 TileQuery Generation

The past work has typically treated a document as a single, monolithic span of text and generated one or more queries to represent the full document. We aim to explore the impact of treating a document as a series of tiles and generating local queries from them to improve retrieval of relevant Wikipedia pages for controversy detection. We call our approach TileQuery as it is based on the TextTiling technique (Hearst, 1997).

We use the block comparison algorithm described by the TextTiling technique to segment a document into multiple paragraphs or "*tiles*". The block comparison method defines a block with k sentences, and computes a lexical similarity score for

every gap between two blocks. When the similarity score dramatically changes at a gap, we assume that is where a sub-topic shift occurs. In this approach, we choose the gaps with the biggest similarity drop between passages as tile breakpoints.

Once we segment the document into tiles, we generate a query that represents each tile. We propose two types of TileQuery depending on whether we treat each tile as a separate document or as part of the document. Note that there are more sophisticated methods for extracting keywords from the given text but that is not the main scope of this work. We aim to compare the effectiveness of a single global query and multiple local queries to retrieve topics of the document in the context of controversy detection, where it is important to retrieve all sub-topics that are covered in the document. To compare the effect of the single global query that are used by Dori-Hacohen and Allan 2013, we use the same query method of selecting top frequent terms. While more sophisticated query generation methods can be applied, this is acceptable in this context because our only goal is to compare the effect of the single vs. multiple local queries.

- **Context-free TileQuery:** Context-free (CF) TILEQUERY takes a view that a document is an aggregation of independent tiles. Each tile is treat is as an independent unit of text and each tile query is generated only within the given tile.

- **Context-aware TileQuery:** Context-aware (CA) TILEQUERY treats each tile as part of a document. A potential issue with the CF-TILEQUERY is that there are some tiles that are hard to understand locally without considering the global context of the document. For example, a document about an author contains multiple tiles on the author's biography, awards, or any excerpt from the author's book. The excerpt should be understood as a context of the author's information, rather than the content of the excerpt itself. In this case, adding the global context helps clarify the topic of each tile, anchoring the tile's query

to the original document. To test this idea, we construct CA-TILEQUERY in the following two ways:

(1) Global/local hybrid Query: This TILEQUERY contains the terms that are selected from each tile as well as the terms that are globally selected from the document. Using the TF query method, we take d most frequent terms from the document, and the $t - d$ most frequent terms from the tile.

(2) Tile Keywords: In this method, tiles are considered as separate documents whereas the document is a collection of those tiles. We compute TF $\cdot$ IDF score among the tiles to find keywords from each tile in the context of the document. TF is considered within the tile, whereas IDF is considered among the tiles of the document. This method, unlike all the other methods, tends to penalize the globally frequent terms throughout the document as they get a low IDF score whereas the terms that are locally frequent within the same tile will be considered to be important keywords.

4.2.3 Aggregating the Ranked Lists

Each TILEQUERY returns a ranked list for relevant Wikipedia pages. We combine these lists to generate a final ranked list for the given document. The intuition behind this aggregation scoring prioritizes documents are ranked high in many tiles. Our scoring function assumes that documents that are retrieved multiple times in several queries and that are ranked high in a ranked list are likely to be more relevant to the overall query document.

$$RelScore(w) = \sum_{l \in R_D} (k - rank_l(w)) \tag{4.1}$$

where k is the number of documents that are retrieved in each ranked list, R_D is a set of ranked lists retrieved from each TILEQUERY of a document query D, and $rank_l(w)$ is the rank of Wikpage w in the ranked list l.

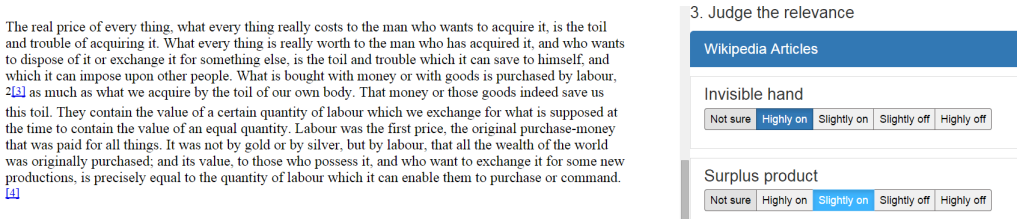

Figure 4.1: An interface snapshot of our annotation website

4.2.4 Intrinsic Evaluation

We first evaluate the query performance on retrieving relevant Wikipages for intrinsic evaluation. We also present the extrinsic evaluation of the query method with regard to controversy detection accuracy in section 4.4.

4.2.4.1 Dataset

To evaluate the performance of TILEQUERY in retrieving relevant Wikipedia topics for a given document, we need an annotated dataset of Wikipedia articles to the query documents. Dori-Hacohen and Allan (2013) previously annotated the relevance of Wikipedia articles to the query documents. For the 377 pages in the controversy dataset, they found the nearest Wikipedia articles using TF10 (i.e., taking the most frequent 10 terms) queries to search engine blekko. For 8,755 unique Wikipedia articles they obtained, they annotated 1,761 articles. We expand this dataset to include more judgments on articles including the ones retrieved by TILEQUERY and ALLQUERY that uses all terms in a document as a query, as another baseline.

For the 377 clueweb documents in the annotated controversy dataset, we generated a candidate set of Wikipedia articles using pooling with TF10, TILEQUERY10 (i.e., taking up to 10 terms for each tile), ALLQUERY (i.e., using all terms in a document). We asked annotators to judge the level of relevance of each Wikipedia article presented in a random order for the given document. Relevance was judged on a five point scale (0 - 4), following the same fashion as Dori-Hacohen and Allan did. We ask how relevant is the given Wikipedia article is to the topic discussed by the Webpage with

the options: "1 - highly **on** topic", "2 - slightly **on** topic", "3 - slightly **off** topic", "4 - highly **off** topic". Figure 7.1 shows an interface of our annotation website where the left panel shows the Webpage content and the right panel shows the list of titles of Wikipedia articles, which are linked to the articles so that annotators can read the content when they are not sure of the relevance. 21 graduate students in Computer Science were recruited as annotators and asked them to judge as many as possible. We obtained 2,248 ratings. For the binarized relevance where the score of 1 and 2 are treated as relevant and 3 and 4 are treated as irrelevant, the judgments show the inter-rater agreement of 74.1%. Out of 303 documents, we obtained at least one judgment rating for 217 documents. Because some documents did not have enough annotations, we evaluated the 132 documents out of 303 that had at least 10 judgments on the binarized relevance.

Table 4.2: The query performance of the three types of TILEQUERY compared to the baseline of TF10 query. * indicates that the difference was statistically significant compared to the baseline.

	MAP	**P@5**	**P@10**	**P@20**
TF10	0.017	0.052	0.041	0.030
CF-TileQuery TF10	0.017	0.061	0.037	0.025
CA-TileQuery TFIDF10	0.008	0.021	0.012	0.009
CA-TileQuery Hybrid 3:7	**0.023**	**0.070***	**0.053***	**0.033**

4.2.4.2 Experiments

We considered three types of TILEQUERY: CF-TILEQUERY, which takes N terms from each tile, two versions of CA-TILEQUERY, one that takes K local terms from each tile and $N-K$ global terms from the document, another one that takes N keywords that have a high TF·IDF score treating tiles as separate documents in a context of the document. As our goal is to investigate the effect of document segmentation in query generation, we similarly use the simple term-statistics-based method such as

TF or TF·IDF as used in the baseline. The average number of tiles within documents was 6.

Table 4.2 shows the query performance of the three types of TILEQUERY compared against the baseline of TF10 method. While the performance of CF-TILEQUERY TF10 and CA-TILEQUERY TFIDF10 performed poorly except for P@5 in CF-TILEQUERY TF10, the results were not statistically significant. The CA-TILEQUERY hybrid query that had global and local terms with 3:7 ratio performed the best, improving 38% in P@5 and 29% in P@10 over the baseline.

This result confirms that our hypothesis that considering the topically-coherent local document text within the global context of the document is more effective in retrieving relevant Wikipedia topics than generating a single query of keywords from multiple subtopics. Adding globally frequent terms to the locally frequent terms helped to keep track of the main topic. CA-TILEQUERY TFIDF10 performed the worst. In that method, since the globally-frequent terms are penalized as they tend to have a low IDF. Among the globally-frequent terms, those who frequently appeared within a tile are more likely to be selected than the the terms that are spread out throughout the document. The result suggests penalizing the globally-frequent terms has the negative effect.

4.3 Solution 2: Smoothing Controversy score of Wikipages

Once the topics of the document are identified in Wikipedia, kNN-WC aggregates the controversy score of the identified topics to estimate the level of controversy of the query document. However, because the existing approaches to estimate the level of controversy are limited in that they rely on dispute signals, the framework is still limited due to the underestimated controversy scores on pages that have not received enough attention.

Due to this phenomenon, even if we retrieve the more relevant topics, if the level of controversy on each topic is erroneous, the final prediction would still be erroneous. Hence, it is necessary to revise these scores to reflect the level of controversy more accurately. If the purpose of the M or C score was to measure the controversy level presented in the Wikipage *per se*, we need newly revised scores that accurately signify controversiality of the topic of the Wikipage *in general*. To do so, we construct a network that connects topically related articles within the Wikipedia. We then revise the controversy score by "smoothing" using the controversy scores of neighbors with more edit history, whose controversy scores can be trusted with a higher confidence.

4.3.1 Constructing a Wikipage Graph with Topically-related Pages

One of the primary reasons why many Wikipages' controversy scores are underrated is that the most controversial discussion has already been delegated in another Wikipage that has a more general topic (Table 4.1). In order to fix the controversy scores of the sub-topical Wikipages, we first construct a tree to identify topically-related neighbors of a Wikipage. Let $G = (V, E)$ be a directed graph where V is a set of nodes and E is a set of edges. In this graph, each node corresponds to a Wikipage and two topically-related Wikipages are connected by a directed edge where edge $e(u, v)$ represents that node v is a sub-topic of u.

As a simple and straightforward yet a high-precision-based method to construct the edges, we consider the pages' titles. If a Wikipage u's title is used as a prefix of other v's title, we assume that u must be a super-topic of v. Because a title is a unique property that each node has and we use nodes' titles to construct edges, we will treat "nodes" and "titles" interchangeably in this context.

To construct a tree for topically-related Wikipages, we define that a node v is a sub-topic of u if v is a child of u, and vice versa. Let the title of v be denoted as T_v an ordered list $[t_1, t_2, ..., t_n]$, where t_i is an i-th space-delimited token of a title. For

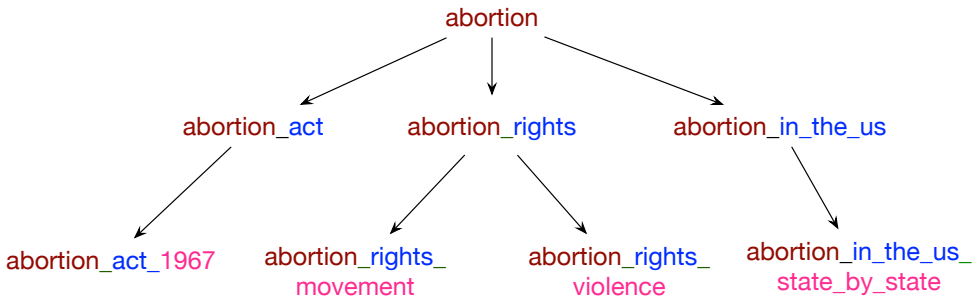

Figure 4.2: An example of the constructed graph for *Abortion* and two different sub-graphs selected based on the two methods. The nodes have more specific titles as they go down from the root as a child node's title has more details added to the current node's title.

example, the title *Abortion act 1967* is represented as [Abortion, act, 1967]. From the list of tokens, we iteratively construct sub-strings by taking the first k terms for $k = 1...n-1$ when n is the number of tokens in the list. The generated sub-strings are sorted in a decreasing order by the length. For example, the generated substrings for the title *Abortion act 1967* would be ["Abortion act", "Abortion"].

While iterating each sub-string from the beginning of the list, the first Wikipage whose title matches to the first sub-string in the list (i.e., the longest sub-string that matches to another Wikipage's title) becomes the direct parent node of this node. For example, when examining "abortion act 1967", the algorithm first encounters "abortion act" as the first longest substring that matches to another page's title. It connects "abortion act" as a parent node of "abortion act 1967". Similarly, "abortion" becomes a parent node of "abortion act" and a grandparent node of "abortion act 1967". Algorithm 1 describes a function to search and construct a parent-child edge for a given Wikipage node.

The graph also contains many noisy relations when the prefix is an ambiguous entity, or a simply too general word, such as "American". To filter out such noisy relations, we only consider that two pages are related if there is a link from one to another in their Wikipage content in addition to this title-relation. Hence, we remove

the edges between two nodes when there is no link between the two Wikipages either any direction. For this, we use publicly available Wikiepdia page-to-page link dataset (Haselgrove, 2009). Figure 4.2 shows an example of constructed graph for the topic of "Abortion". From the filtered graph, we finally revise the controversy score using smoothing.

Algorithm 1 An algorithm for finding parent nodes for a given Wikipage node

```
1: procedure findTopicParent(v, V)                ▷ Find parent nodes for v in V
2:     parents = set()
3:     l = list(tokenize(v.title))
4:     n = len(l)
5:     for i = n-1 to 1 do
6:         titleSubstr ← concatenate(l, 0, i)
7:         for all w ∈ V do
8:             if titleSubstr = w.title then
9:                 v.parent = w
10:                w.child.add(v)
11:                return
12:            end if
13:        end for
14:    end for
15: end procedure
16:
17: procedure constructTree(V)
18:    for all w ∈ V do
19:        findTopicalParents(w, V)
20:    end for
21:    return V
22: end procedure
```

4.3.2 Graph-based Smoothing

When a Wikipage is given as a query, we extract a sub-graph around the node from the constructed graph using one of the two methods, whose examples are demonstrated in Figure 4.2:

- **Direct Family**: A sub-graph around the query node including its children and its parent. The resultant graph only consists of nodes that have a direct prefix-contain relation with the query node.

- **Extended Family**: In addition to the sub-graph obtained by the above method, sibling nodes that share the same parent with the query node are added. Although siblings may not be topically related to the query node especially if the parent (i.e., prefix) is a general term, this allows broader coverage of potentially related pages.

Once we obtain the sub-graph, we treat all nodes in the sub-graph as topically-related neighbors of the query node. Using these topically-related neighbors, we perform smoothing on each node's controversy score. For smoothing, we assume that the controversy score of a Wikipage with more revision history is more reliable. For the query node w, We first obtain a weighted sum between w and a neighbor node v based on their "reliablitily", which is computed from the ratio of their revision counts. The smoothed scores between w and other neighbors are aggregated via another weighted sum based on how reliable each neighbor is. Given an original controversy score $c(w)$ of a node w, the smoothed score $c'(w)$ is computed as follows:

$$c'(w) = \sum_{w_i \in \mathcal{N}(w)} f(r_i) \cdot \left(c(w) \cdot (\frac{r}{r + r_i}) + c(w_i) \cdot (\frac{r_i}{r + r_i}) \right) \tag{4.2}$$

where w is a given Wikipage, $c(w)$ is a controversy score of w, r_i is a revision count of w_i, $\mathcal{N}(w)$ is a set of neighbor Wikipages of w, $f(r_i)$ is a fraction of r_i among the revision counts of neighbors and computed as $\frac{r_i}{z}$ and $z = \sum_{w_k \in \mathcal{N}(w)} r_k$.

4.3.3 Aggregation and Voting

We summarize the aggregation and voting schemes introduced by previous work. Once the controversy scores are obtained for k Wikipages, we aggregate the k scores

Table 4.3: An example of two controversy scores on several Wikipages on "Abortion", before and after score smoothing

	Original scores		Revised scores		Revision Count
	M	C	M	C	
Abortion	4,102,593	0.300	3,718,532	0.269	12,300
Abortion_Act_1967	0	0	1,966,410	0.146	168
Abortion_in_Canada	0	0	1,906,494	0.142	942
Abortion_in_the_United_States	0	0.002	1,828,736	0.135	2,281
Abortion_Law	0	0.003	1,877,349	0.139	1,387

Table 4.4: Accuracy, F1, and the best parameters in 5-fold runs for different query and inferred score settings.

ID	**Query**	**Smoothing**	**K**	**C Threshold**	**M Threshold**	**Aggregation**	**Acc.**	**F1**
1		None	{5, 20}	$4.18 \cdot 10^{-2}$	20000	{M, Maj.}	0.731^{47}	0.678
2	ALL	Direct	15	$4.18 \cdot 10^{-2}$	{84930,20000}	{M, Maj.}	0.760^{14}	0.679
3		Extended	{5, 20}	$4.18 \cdot 10^{-2}$	{20000, 84930}	{M, Maj.}	0.764^{14}	0.675
4		None	20	$0.17, 4.18 \cdot 10^{-2}$	{20000, 40000}	{M, Maj.}	0.720	0.575
5	TF10	Direct	20	$4.18 \cdot 10^{-2}$	{20000, 84930}	{M, Maj.}	0.757^{14}	0.680
6		Extended	{10, 20}	$4.18 \cdot 10^{-2}$	{20000, 84930}	{M, Maj.}	0.761^{14}	0.678
7		None	{10,15,20}	$4.18 \cdot 10^{-2}$	{20000, 84930}	{M, Maj.}	0.723^{4}	0.635
8	TILE	Direct	20	$4.18 \cdot 10^{-2}$	20000	M	0.812^{146}	0.766
9		Extended	{10,15,20}	$4.18 \cdot 10^{-2}$	20000	{M, Maj.}	0.796^{23467}	0.745

by taking the average or max of them. Since we use three different scores, M, C, and D, three aggregated scores, M_{agg}, C_{agg}, and D_{agg} are computed. We turn these scores into binary label indicating controversial (1) or non-controversial (0), using corresponding thresholds. $M_{label} = 1$ if $M_{agg} \geq Threshold_M$, and 0 otherwise. Using the three generated labels, we use a voting scheme to make a final decision. We test 6 voting schemes as parameters in our experiments.

The webpage is controversial if:

— **C/M/D**: $\{C_{label}, M_{label}, D_{label}\}$ is 1, respectively.

— **Majority**: the majority (i.e., at least two) of $\{C_{label}, M_{label}, D_{label}\}$ is 1.

— **Or/And**: $C_{label}\{\vee/\wedge\}M_{label}$ $\{\vee/\wedge\}D_{label}$ is 1.

4.4 Experiments

4.4.1 Dataset

We use the publicly available controversy dataset[1] released by Dori-Hacohen and Allan (2013). The dataset consists of 303 webpages from the ClueWeb09 collection, which is a publicly available dataset of crawled general webpages (Callan and Hoy, 2009). Note that the annotated webpages do not include any Wikipages. Each document is annotated with the controversy level of four scales: 1 - "clearly controversial", 2 - "possibly controversial", 3 - "possibly non-controversial", and 4 - "clearly non-controversial". To convert the annotations to binary judgments, we treated the documents with average ratings among annotators of less than 2.5 as controversial, and otherwise non-controversial as done by the previous work (Dori-Hacohen and Allan, 2015). Of 303 documents, 42% of them are labeled as controversial. For retrieving Wikipedia pages as topics, we leverage the Wikipedia dump of 2013-Jun-05.

To extract queries from the actual content of a webpage, we remove peripheral text that specifies layout (e.g., HTML, CSS, and JavaScript) and so-called "boilerplate" material (e.g., navigation links, advertisements, headers, and footers). Leaving these material in the document leads to over-representation of several non-content words and phrases, such as "home" in the menu, or "all rights reserved" in the footer, that otherwise might cause noisy terms to be included in a query. We removed this non-content information using the open source library jusText[2].

4.4.2 Experiment Setup

To test the effectiveness of TILEQUERY and controversy smoothing, we consider two other query methods as the baselines. One is TF10, the 10 most frequent terms, as in the prior work. As taking only k terms as in a query might miss information,

[1] http://ciir.cs.umass.edu/downloads

[2] https://code.google.com/p/justext/

Table 4.5: Improvements of accuracy and F1 score between runs and their statistical significance tests

Run 1	Run 2 (Better)	$\lvert Acc_1\text{-}Acc_2\rvert$	$\lvert F1_1\text{-}F1_2\rvert$	p value	Significant?
All & None	All & Smoothing (D)	2.9%	0.1%	0.0003	*
All & None	All & Smoothing (E)	3.3%	0.3%	4.11e-05	*
All & None	TF & None	1.1%	10.3%	1.54e-10	*
All & Smoothing (E)	TF & None	4.9%	10.0%	0.0017	*
All & Smoothing (E)	TF & Smoothing (D)	0.7%	0.5%	0.0889	*
TF & None	TF & Smoothing (E)	4.1%	10.3%	1.96e-05	*
TF & None	Tile & None	0.3%	6.0%	1.96e-05	*
Tile & None	All & None	0.8%	4.3%	0.0035	*
TF & Smoothing (D)	Tile & Smoothing (D)	5.5%	8.6%	0.0909	
Tile & Smoothing (D)	Tile & Smoothing (E)	1.6%	2.1%	1.0000	

we consider another baseline, ALL query that uses all terms in a document as a query to observe the extreme case of TFN. Therefore, we have three query methods – TF10, AllQuery, tilequery – and three score smoothing setup – None (baseline), smoothing with a direct family (D), smoothing with an extended family (E) –. Finally, we consider all 9 pairwise setting of three query methods and three score smoothing setups (Table 4.4).

In each setting, we varied the four sets of parameters, the number of neighbors K (1, 5, 10, 15, 20), aggregation method (avg, max), voting methods (C, M, D, Majority, Or, And, D$\vee$(C $\wedge$ M)), and thresholds for C and M as tested in the prior work. Run 4 is the setting proposed in the prior work (Dori-Hacohen and Allan, 2015). We found the best parameter setting for each run using 5-fold cross validation with the target metric accuracy. Thus, for the 9 settings, there are 5 sets of parameters learned for each fold. We used McNemar's Test (1947) for statistical significance test.

4.4.3 Results and Discussion

We present our experimental results in Table 4.4 and its statistical significance test results in Table 4.5. When there is no smoothing on the controversy scores, among the three query methods considered – AllQuery, TF10, tilequery –, AllQuery showed the highest performance both in accuracy and F1 score, followed by tile-

QUERY (run 1, 4, 7 in Table 4.4). In all settings, using controversy score smoothing significantly improved the classification accuracy and F1 score. In fact, runs with smoothing outperforms runs with any of the query method without smoothing. For example, the runs with any type of smoothing (run 2,3,5,6,8,9) show higher performance than the run 1 of ALLQUERY, the best query method without smoothing. While without smoothing ALLQUERY performed the best, TILEQUERY is shown to be most effective with any smoothing combined. Between using two types of smoothing of a direct and an extended family, the "extended family" performed better with ALLQUERY and TF10 while the "direct family" performed better with TILEQUERY. However, our statistical significance test suggests that the differences are not statistically significant.

4.5 Conclusion

In this chapter, we revisited two assumptions of the kNN-WC model: Based on the derived model in Eq.4.3, the success of the algorithm depends on how accurately the two probabilistic components are being estimated: $P(w|q_D)$, the probability that a given Wikipage is a relevant topic to the query q_D and $P(\mathit{contention}|w)$, the probability that a retrieved Wikipage shows a high contention among the Wikipedia editors.

$$P(c|D) = \sum_{w \in \mathcal{W}_D} \big[P(\mathit{contention}|w; \Omega_w) \cdot P(w|q_D)\big] \tag{4.3}$$

We revisit the kNN-WC algorithm, a specific implementation proposed by Dori-Hacohen and Allan (2015). We point out that the algorithm could be improved to better implement the model by ensuring that the algorithm satisfies the two assumptions more accurately. We recap the two assumptions here and how we addressed to satisfy the assumptions better.

A1: $P(w|q_D)$ assumes that a query generated from the document retrieves Wikipages that represent the document's topic.

To generate more effective query to retrieve the relevant Wikipedia topics, we have proposed a new query method named TILEQUERY that extracts multiple queries from topically-coherent paragraphs in a document.

> **A2:** $P(contention|w;\Omega_w)$ **assumes that Wikipages that discuss controversial topics will show a high level of contention among the editors of the page, and vice versa.**

We have observed that $P(contention|w)$ that is estimated from existing Wikipedia controversy scores is often inaccurate and underrated for Wikipages that did not receive enough attention, or whose controversial discussion has been delegated in another page with a broader topic. We proposed a modification to the existing Wikipedia controversy scores to infer more accurate and reliable scores via smoothing using topically-related neighbors in Wikiepdia. From our experiments, the effect of the controversy smoothing alone seems to be more significant than the effect of a query method alone. Using the proposed query method along with the smoothing showed the best performance, increasing the accuracy by 9% and F1-score by 19% points.

However, we would like to point out that this issue stems from not just the implementation choice, but from the inherent property of the model to some extent. We previously stated via P1 that kNN-WC model is designed to be bounded by the potential limitations of the models of $P(c|w)$. The kNN-WC model calls a population-based topic-controversy model as sub-component, which require evidence of disputes for the given topic instance. These models tend to have a high precision but suffer from relatively low recall. They are good for analyzing the given controversial signals but tend to make false judgments when the contentious signals are not present. It is hard to distinguish the cases where the topic is not controversial or controversial but simply missing the signals of contention for the moment or for that topic instance. Therefore, we address this issue in Chapter 6 by estimating the controversy score of topics that change over time beyond the observed signals. Lastly, we would like

to stress that any implementation of the kNN-WC model should take this issue into consideration to expect a good performance in a real dataset.

CHAPTER 5

CONTROVERSY LANGUAGE MODELS

5.1 Counter Properties for the New Model

We have identified three properties that kNN-WC model has in Chapter 3.2. While we have proposed modifications for a better implementation of kNN-WC model to improve the empirical performance, the proposed algorithm is still likely to be bounded by the assumptions and the properties of the underlying model. In this chapter, we propose a new model that challenges these properties in the pursuit of an approach that has complementary characteristics to kNN-WC model. We recap the three properties of kNN-WC model and propose the counter property that the new model should have by challenging each property:

> **P1: The model has a population-based topic controversy model as a sub-component.**
>
> **P1′: A model does not depend on explicit "contention" signals that are generated from people's reactions and behaviors.**

Due to P1, kNN-WC model is inherently limited in efficacy and adaptability because "contention" signals such as disputes are expensive signals because they require people to engage in the discussion or to show reactions. In addition, the presence of "contention" signals are easily delayed until enough people participate and generate a contentious discussion, if they do, ever. We have shown in Chapter 4 that the "contention" signal is not reliable because it is selectively available, which resulted in many Wikipages whose topics are controversial but do not contain such signals. Hence, for the new model, we consider an alternative property that the model does

not depend on the population-based contention signals to estimate the probability of controversiality. We do so by transferring "contention" signals to "language" features.

P2: Non-controversiality is not directly modeled.

P2′: Non-controversiality is explicitly considered for the classification of a document's controversiality.

The kNN-WC model does not directly consider the probability that a document is non-controversial. This means that when a document contains more non-controversial keywords, it does not directly decrease the probability of controversy because the probability of controversy is more affected by the presence of controversial keywords. Instead of defining non-controversiality simply as a lack of controversy signals, we consider a counter property of explicitly considering non-controversiality of the document for the final decision. For example, the new property assumes that the document is controversial if the controversial content is dominant compared to the non-controversial content.

P3: A document's text is only a proxy to find topics.

P3′: A document's text is directly considered for estimating the probability of controversiality.

Instead of only using the documents' text as a proxy to find topics, alternatively, we propose to directly the documents' text to estimate the probability of controversy. While this original property of kNN-WC model is likely to yield the same probability of controversy for the two documents once they retrieve the same controversial topics, the alternative property of the new model will allow to distinguish if one document is more controversial by considering the language of the original text.

5.2 Proposed Model

Therefore, we explore another probabilistic model of controversy to satisfy the new counter properties. We aim to use an alternative "language" signal and also directly model non-controversiality for the controversy classification. Lastly, we directly

consider the language of the document's text to estimate the probability of controversiality. As part of our effort to find a new model for controversy detection, we first turn to social science research to understand how controversy is being identified and shaped.

The most relevant work to our interests would be Cramer's (2011). Cramer explains that "controversy" cannot necessarily be verified to exist in the world independent of its appearance in text, but rather it is created and shaped by the discourse surrounding it, particularly in news outlets. He refrains from defining the term directly, referring to it as a "metadiscursive" (terms that are used to denote a discussion of discussion) and "indexical" (terms whose specific meaning changes from context to context) term, meaning that it may be difficult to formulate a mathematical or technical definition of controversy, and it can be loosely defined as *something that you would know when you see it*. However, Cramer's work suggests that language could be a key feature in identifying controversy.

Cramer manually studies patterns of text surrounding specific terms such as `controversy`, `dispute`, `scandal`, and `saga` within the Reuters corpus (Rose and Whitehead, 2002), as being indicative of controversy. Motivated by Cramer's research, we propose a new probabilistic model of controversy that considers how similar the document's language is to the one that discusses a range of controversial topics.

Table 5.1: The notation summary of controversy language model

Symbol	Meaning
L_C	A language model of controversy
L_{NC}	A language model of non-controversy
L_G	A background language model of all topics
D_C	A set of controversial documents used to build L_C
D_{NC}	A set of controversial documents used to build L_{NC}
$tf(w, D)$	The frequency of term w in a document D
$P(w\|L)$	The probability of term w in the language model L

We defined that $P(c|D)$ indicate the probability that D is controversial and $P(nc|D)$ the probability that D is non-controversial. We set $P(c|D) + P(nc|D) = 1$ in Chapter 3.3. In this model, we classify that the document is controversial if $P(c|D) > P(nc|D)$ holds. The idea behind this assumption is that the controversiality of the document should dominate the non-controversiality of the document to be classified as "controversial." Because we are only interested in whether $P(c|D) > P(nc|D)$ holds rather than the actual probabilities, so we can use rank-safe approximations.

Each of $P(c|D)$ and $P(nc|D)$ can be represented using Bayes' theorem, which allows us to consider the following odds-ratio:

$$\frac{P(c|D)}{P(nc|D)} = \frac{P(D|c)}{P(D|nc)} \cdot \frac{P(c)}{P(nc)} > 1 \tag{5.1}$$

Now our test condition can be expressed as:

$$\frac{P(D|c)}{P(D|nc)} > \frac{P(nc)}{P(c)} \tag{5.2}$$

where for our purposes, we can treat the right hand side as a constant threshold (since it is independent of the document D), which can be learned with training data. To avoid underflow, we actually calculate the log of this ratio. The higher this log-odd score is, the more distinctively a given term appears in the controversial topic corpus than in the non-controversial topic corpus:

$$\log P(D|c) - \log P(D|nc) > \alpha \tag{5.3}$$

Therefore, we only have to estimate the probabilities $P(D|c)$ and $P(D|nc)$, which we do using the language modeling framework by the construction of a language model of controversy L_C, and a non-controversial language model L_{NC}. We make the standard term independence assumption for each word (v) in our document (D), and

avoid zero probabilities with linear smoothing. We create another language model L_G for the purpose of smoothing using a broad "background" collection of documents, as opposed to controversial and non-controversial collections. In practice, we estimate both the general language model (L_G) and the non-controversial language model (L_{NC}) as the same by constructing them from the set of all documents.

$$P(D|c) \approx P(D|L_C) = \prod_{v \in D} (\lambda P(v|L_C) + (1-\lambda) P(v|L_G)) \tag{5.4}$$

$$P(D|nc) \approx P(D|L_{NC}) \approx P(D|L_G) = \prod_{w \in D} P(v|L_G) \tag{5.5}$$

Here, D_C is a set of controversial documents, and D_{NC} is a set of non-controversial documents, which we estimate in our collections as the background collection, D_{BG}.

$$P(w|L_C) = \frac{\sum_{d \in D_C} tf(w,d)}{\sum_{d \in D_C} |d|}, P(w|L_{NC}) = \frac{\sum_{d \in D_{BG}} tf(w,d)}{\sum_{d \in D_{BG}} |d|} \tag{5.6}$$

where $tf(w,d)$ in dicates the term frequency of w in d and $|d|$ is the length of d. Therefore, to build a language model of controversy, we need to find D_C. We explore Wikipedia Controversy Features (WCF) and Cramer-inspired query based models to construct D_C as following:

- **Highly Contentious Articles** While the normalized WCF features are used to estimate $P(Contention|w;\Omega_w)$ in kNN-WC model, we simply take the top K articles that have high WCF values in Wikipedia. In our experiments, three types of WCF, M/C/D scores are considered.

- **Controversy-indicative terms:** Documents that are retrieved by a query believed to indicate controversy. We explore Cramer's terms as well as manual lexicons from past work (Mejova et al., 2014; Roitman et al., 2016). The examples of these terms is shown in Table 5.2.

Table 5.2: An example of controversy-indicative terms.

Reference	Search Terms
Roitman et al.	dispute, disputable, disagreement, debate, polemic, feud, question, schism wrangle, controversy, dispeace, dissension, criticism, argue, disagree, claim argument, conflict, opposition, adversary, antagonism, oppose, object, case, loggerheads, quarrel, fuss, moot, hassle, altercate, evidence, clash, issue, problem, emphasize, recommend, suggest, assert, defend, maintain, reject, support, challenge, doubt, refute, confirm, prove, validate, establish, concur substantiate, verify, against, resist, support, agree, consent, accept, refuse plead, right, justify, justification
Mejova et al.	abuse, administration, afghanistan, aid, american, army, attack, authority, ban, banks, benefits, bill, border, budget, campaign, candidate, catholic china, church, concerns, congress, conservative, control, country, court, crime, crisis, cuts, debate, debt, defense, deficit, democrats, disease, dollar, drug, economy, education, egypt, election, enforcement, fighting, finance, fiscal, force, funding, gas, government, gun, health, immigration, ...
Cramer et al.	controversy, dispute, saga, scandal

5.3 Evaluation

We leverage the same controversy dataset introduced in Chapter 4 that consists of judgments for 303 webpages. We perform 5-fold cross-validation and report measures on the reconstructed test set.

We implement the kNN-WC model as the baseline, both the original algorithm and the improved version of it introduced in Chapter 3. In order to construct D_C, we needed the text of Wikipedia itself. Unfortunately, obtaining the same version of dumps as those used in prior work (Das et al., 2013; Dori-Hacohen and Allan, 2015; Yasseri et al., 2012) is nearly impossible. For ease of future reproducibility, we leverage the long abstracts from the 2015-04 release of DBPedia (Lehmann et al., 2015)

Prior work reported accuracy; we note that 65% of the 303 documents were non-controversial, so that accuracy does not provide the best view of this dataset. In this work, we primarily present results using the Area Under the Curve (AUC) measure, as we can compare performance without tuning thresholds. While AP and MAP have the same advantage for not requiring a threshold, AP explicitly gives advantages

Table 5.3: The accuracy of the models.

Models	Accuracy
The kNN-WC algorithm (Dori-Hacohen and Allan, 2015)	0.737
The improved kNN-WC algorithm (Chapter 4)	0.796
CLM	0.779

Table 5.4: Wikipedia-Based Controversy Detection Approaches. All Controversy Language Model (CLM) approaches have significant improvements over their respective kNN-WC counterpart at the $p < 0.05$ level.

Method	WCF	AUC
kNN-WC model	M	0.733
kNN-WC model	C	0.743
kNN-WC model	D	0.500†
CLM	M	0.801
CLM	C	0.835
CLM	D	0.795

† In the kNN-WC-D approach, no neighbors were found with dispute tags, so it is equivalent to the weak baseline performance of the *NO* classifier.

to a method that correctly predicts a few top-ranked items, which makes it mores suitable for Information Retrieval tasks rather than classification tasks like ours (Su et al., 2015). Since accuracy was used in prior work, we report it as well in Table 5.3: Compared to kNN-WC algorithm, we improve from 0.72 accuracy (as reported by Dori-Hacohen and Allan (2015) and 0.737 accuracy (as reproduced) to 0.779 ($p < 0.001$). We also report the accuracy of the improved version of the kNN-WC algorithm proposed in Chapter 4. For our statistical significance tests, we follow in the footsteps of the `pROC` (Robin et al., 2014), and obtain confidence intervals from bootstrap resamples of the predictions.

For each fold, we trained two parameters by grid search: K, the number of top documents to choose, and λ, the smoothing parameter. For example, to create our M-score-based language model, we ranked the documents in our Wikipedia collection by their M score, and derived a language model based on the concatenation of the top K documents. These models are presented in Table 5.4.

Table 5.5: Language Models built from documents relevant to Cramer's controversial terms (Cramer, 2011). Collection size $|C|$ in millions of documents and type is shown for comparison of results. We found that our wiki dataset was significantly better than all others, which had no pairwise differences otherwise.

Expansion Dataset	Type	$\|C\|$	AUC
DBPedia	Wiki	4.6M	0.853
ClueWeb09B (Spam60)	Web	33.8M	0.741
Reuters	News	0.8M	0.745
NYT-LDC	News	1.8M	0.710
Robust04	News	0.5M	0.711
Signal-1M	News	1M	0.710

Table 5.6: Language Models built from Cramer's terms and existing lexicons on DBPedia. We find that "controversy" is the most indicative term, and that "saga" is no better than random. Combining terms led to no improvement over "controversy" alone.

Query to build D_C	AUC
`controversy`	0.856
Roitman (Roitman et al., 2016)	0.823
`dispute`	0.740
`scandal`	0.721
Mejova (Mejova et al., 2014)	0.698
`saga`	0.500

For building Cramer language models, where the relevant document sets were not created by WCF, we used the Galago search engine to rank documents using a query-likelihood retrieval. We explore 6 different corpora as document sources (Table 5.5). The K highest-scoring documents were then used as our controversial document set: D_C.

5.4 Results

In Table 5.4, we present results of our models built around WCF. All our language modeling approaches are significantly stronger than the k-NN derived approaches. We only report results of WCF features independently because methods of aggregating

Table 5.7: A comparison of lexicons built manually and through crowd-sourcing in prior work to our automatically derived language models. A (*) indicates significant improvement over the best lexicon approach. "TF10" indicates that the TF10 query is used to represent a document whereas "Full" indicates that the full text of the document is used as a query.

Method	Document Query	AUC
Roitman Lexicon (Roitman et al., 2016)	TF10	0.543
Mejova Lexicon (Mejova et al., 2014)	TF10	0.562
Mejova Lexicon (Mejova et al., 2014)	Full	0.615
Roitman Lexicon (Roitman et al., 2016)	Full	0.695
Cramer Language Model	Full	0.783
WCF Language Model	Full	0.823*
WCF Language Model	TF10	0.835*
Cramer Language Model	TF10	0.856*

these features did not improve significantly over the best feature, and these methods were not quite comparable across kNN-WC and LM approaches.

In Table 5.5, we present an initial exploration of Cramer's hypothesis that news is able to name and define controversy. While we were pleasantly surprised by the efficacy of this simple approach, we did not see the best performance in the news corpora (Rose and Whitehead, 2002) used by Cramer, but rather in using DBPedia as the expansion set. We also explored this approach on other news datasets (Robust04, NYT-LDC (Sandhaus, 2008), and Signal1M (Corney et al., 2016) but results were statistically equivalent on all news corpora we tried. Attempting to correct for the fact that some news corpora are no longer modern, we explored the contemporary Signal Media News Dataset (Corney et al., 2016), and attempting to correct for the size differences in the better-performing corpora (DBPedia (Auer et al., 2007) and ClueWeb), we explored the larger NYT-LDC corpus (Sandhaus, 2008).

While Cramer defined four keywords to be indicative of controversy, we find that "`controversy`" dominates effectiveness on this dataset. We explore these keywords as queries into an expansion corpus, and construct a language model from the highest scoring documents for the given query. That language model is then used for classi-

Figure 5.1: The top controversial terms of CLM that have a high log odds score (Eq. 5.3) and are frequent in the corpus. Note that the size of the font is a layout choice and does not mean that the term has a higher probability. Colors of the text are chosen arbitrary.

Figure 5.2: The top controversial terms of CLM that have a high log odds score (Eq. 5.3) and are frequent in the corpus. Colors of the text are chosen arbitrary.

fication. Mejova et al. (2014) and Roitman et al. (2016) presented manually-curated lexicons for controversy tasks. We explore their use intrinsically, with Jaccard Sim-

ilarity between the lexicon and the document terms in Table 5.7 and as queries to build a language model in Table 5.6.

Lastly, to understand the characteristics of the model, we extract the top representative controversial terms and non-controversial terms in CLM. Because the top terms that have the high log-odd scores (Eq. 5.3) are often extremely rare terms (e.g., rare terms that only occurred in the controversial corpus but not in the non-controversial corpus at all), we also weighted the terms by its frequency multiplied by the log-odd score for the presentation in Figure 5.1 and 5.2. While the "controversy-indicative terms" proposed from past work contain metadiscursive terms that signal disputes such as "dispute", "disputable", "refuse" (refer to Table 5.2), the terms from CLM are mostly topical. The top controversial terms of CLM include topical terms such as "homemopathy", "falun gong", "jehovahs", "anarchism", whereas the non-controversial terms tend to have broader topics such as "university", "company", "family", and "albums".

As the controversy test dataset is relatively small, we were concerned about the possibility that the controversy document collection used for building CLM happen to include all of the specific controversial topics appeared in our test set. The best run from CLM was built with DBPedia using the query "`controversy`". As the best run used the top 241 documents, we examined those documents to look at the overlap between the train and test collection (see Appendix A). The list contained a lot of specific controversy cases unlike the list from high M scores. Several controversial topics in the test set documents such as "creationism", "homeopathy", and "capitalism" were not included in the training corpus, but CLM was still able to identify controversy from those documents.

Table 5.8: The ratio of the documents that are correctly and incorrectly classified by kNN-WC and CLM).

Controversial	**Correct by KNN**	**Wrong by KNN**
Correct by CLM	**69 (65%)**	**8 (8%)**
Wrong by CLM	**7 (6%)**	**22 (21%)**
Non-controversial	**Correct by KNN**	**Wrong by KNN**
Correct by CLM	**153 (78%)**	**8 (4%)**
Wrong by CLM	**13 (7%)**	**23 (12%)**

5.5 A Comparison Between kNN-WC and CLM

To understand the different characteristics of the two approaches, we examine the cases where one makes a correct classification and the other does not, and vice versa. The kNN-WC algorithm made slightly more errors than CLM for classifying controversial documents with the mis-classification rate of 8% for the kNN-WC algorithm and 6% for CLM. On the other hand, CLM made more errors than the kNN-WC algorithm for classifying non-controversial documents with the mis-classification rate of 7% for CLM and 4% for the kNN-WC algorithm. This suggests that kNN-WC algorithm is slightly more prone to make false negative errors whereas CLM is more prone to make false positive errors.

We observed the distribution of the document length of the documents (i.e., the number of terms) that are labeled as controversial and non-controversial by each method to see how the document length affects each method's classification decision. Figure 5.3 and 5.4 show the distributions of the document length that are classified by each method for controversial and non-controversial documents. Shorter documents tend to be classified as controversial more often by CLM whereas the kNN-WC algorithm has the opposite tendency compared to the human labels.

We manually analyzed the cases of the documents that were correctly classified by kNN-WC while being incorrectly classified by CLM and vice versa to understand the reasons for mis-classifications. In the kNN-WC model, because the controversy

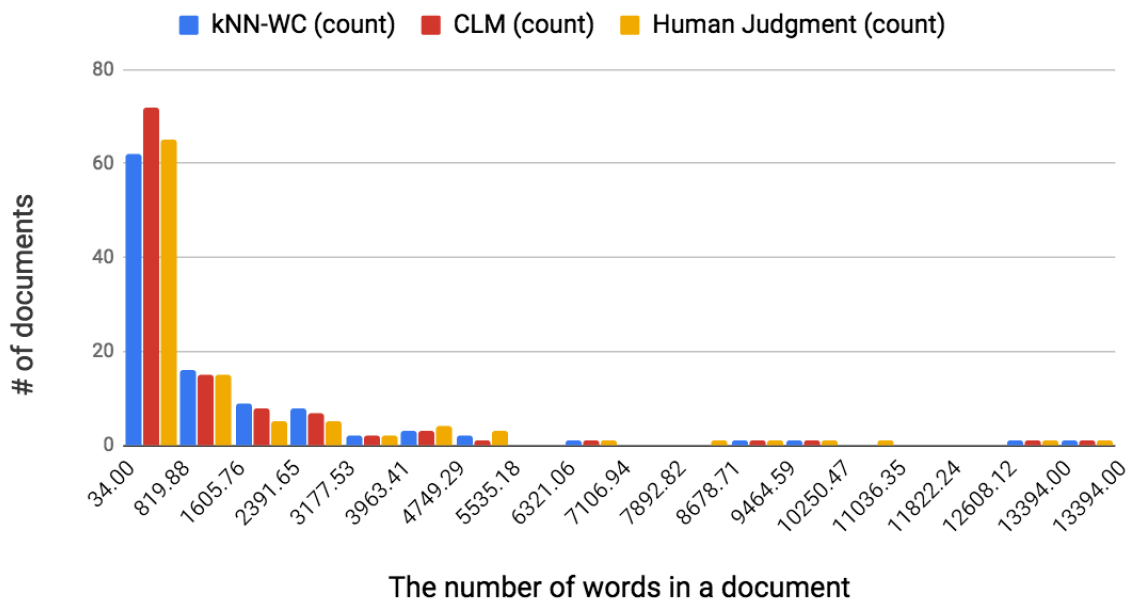

Figure 5.3: A distribution of the document length of documents that are labeled as **controversial**

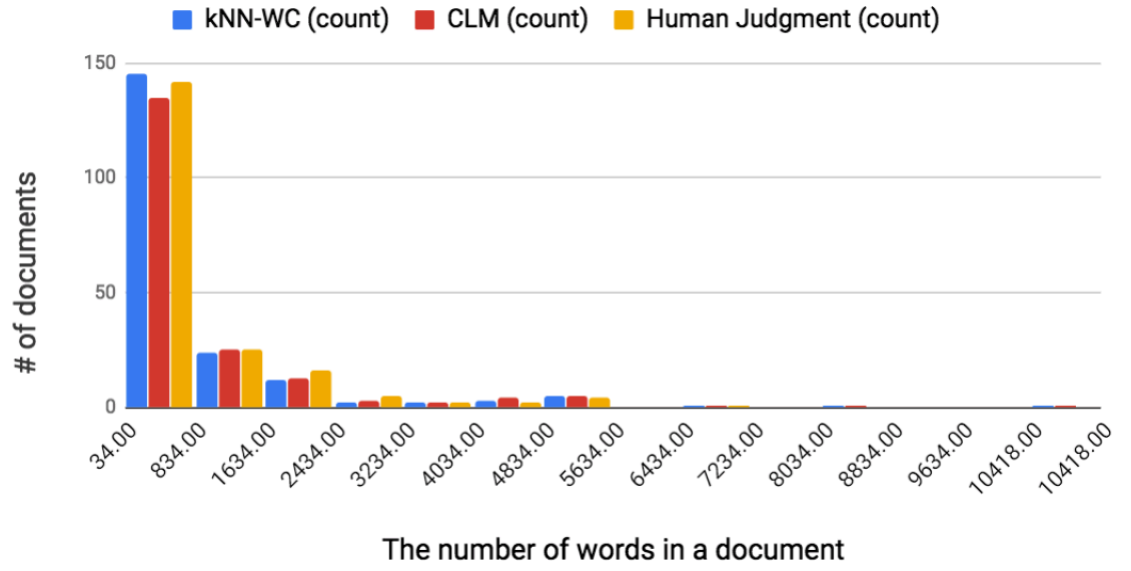

Figure 5.4: A distribution of the document length of documents that are labeled as **non-controversial**

judgment is estimated from the retrieved Wikipages from the document query, the mention of controversial keywords in the document has an indirect effect in the final

Table 5.9: The top 10 log-odd score terms of four documents as well as their gold standard label and CLM labels.

Document ID	Top 10 log-odd score terms	Gold label	CLM label
clueweb09-en0008-16-31383	analogy creationist intelligent crucify naturalism evolutionary evolution argument objection debate	C	C
clueweb09-en0000-47-35377	monotheistic devotions analogy mecca hadith quran racial prophet muhammad tenet	NC	C
clueweb09-en0011-89-02679	homeopathy people speak speaker 2009 raise running friends back june	C	NC
clueweb09-en0007-51-03335	editorial including resources mention bring recording any the to com	NC	NC

judgment. However, once the controversial topic of Wikipage is retrieved, highly controversial topics tend to dominate the probability of document's controversiality. Once a highly controversial topic is retrieved in the list, no other non-controversial topics in the list can cancel it out. On the other hand, in CLM, the mention of a certain controversial keyword might not be likely to dominate the probability of controversiality. However, each mention of the controversy keyword directly affects the document's probability of being controversial in CLM.

While in kNN-WC, the effect of controversy keywords is diluted because the level of controversy is measured from the retrieved topics from the query, whereas in CLM, the effect of having controversy keywords is more direct. However, the impact of retrieving controversial topics is more influential in kNN-WC model than in CLM. This suggests that kNN-WC implements the principle that as long as the document discusses a controversial topic, no matter how much it also discusses non-controversial topics, it should still be classified controversial.

Table 5.11 shows an example of a highly controversial document that argues that abortion is a cause of breast cancer. The document was correctly labeled as "controversial" by kNN-WC while being inaccurately labeled by CLM. In kNN-WC, the top-ranked topic "Abortion-breast cancer hypothesis" was highly relevant to the content of the document. The original M score and C score of this topic was 1550 and 0, which is considered to be non-controversial by the threshold of the algorithm. The

smoothing method introduced in Chapter 4 corrected the scores to be 178,961 and 0.0131, respectively. The Wikipage ranked at 10, "Abortion" that has a high M and C score, also helps to classify this page to be controversial. Being able to retrieve specifically relevant topics such as "Abortion-breast cancer hypothesis" is one of the biggest advantages of kNN-WC model, which comes from the benefits of a general k-nearest-neighbor model.

In kNN-WC, the presence of the highly controversial topic "Abortion-breast cancer hypothesis" and "Abortion" in the ranked list, which had a very high M and C score, often dominantly determines the document to be controversial, as either using the average or max aggregator of the retrieved scores, it results in a highly controversial score. However, in CLM, while the terms such as "abortion" and "pregnancy" had a high probability of controversy, the decision is usually made by considering other factors. Having more non-controversial terms may cancel out the controversiality of the document in CLM.

Table 5.9 shows another example of four documents with their top 10 log-odd score terms as well as their gold standard and CLM labels. While for the two cases where the gold labels and CLM labels match, the extracted terms reasonably contained the controversial and non-controversial keywords. For the other two cases where the labels do not match, they illustrates the situation where the topic of the document was controversial, but the document did not particularly say anything controversial. For example, document 'clueweb09-en0011-89-02679' contains an advertising text for their homeopathy-related events. While the topic of homeopathy itself is controversial, the annotator decided that the document does not contain any controversial content.

5.6 Limitations

While CLM is constructed from the language of controversial topics, it is obviously not aware of newly-emerged keywords or the controversial entities that did not exist

in the training corpus. From our analysis, when a new controversy arises, CLM is still able to catch that there is some controversial event because even the new controversy tends to include keywords that are highly correlated to any controversial event. For example, during the Facebook–Cambridge Analytica scandal, another controversy arose when an internal memo by Facebook Vice President Andrew "Boz" Bosworth that was criticized for justifying "bullying" and "terrorism" at the cost of the company's growth (Ryan Mac, 2019). When we analyze the tweets of the given day using CLM built from Wikipedia's top controversial articles, the model fails to capture "Andrew Bosworth" as a controversial entity, while it still captures "leaked" or "terrorism" as controversial keywords. For other new controversies, the similar pattern occurs. We believe that CLM is still able to capture the new controversies that were not included in the model, but without "understanding" the actual controversial topic. However, for the same reason, CLM is susceptible to make false positive errors. The model also inherently suffers from the fact that it is a global model that combines all controversial topics. This can be allevated by building a domain-specific or a query-specific, time-adaptive CLM, which we leave it as future work.

5.7 Conclusion

We challenge the three properties presented from the previous work and propose a new model that complements them. Using insights from recent social science research, we motivate and explore the first language modeling approach to detecting controversy. We find that our new approach is statistically better than prior work, while being simpler. We explore strongly controversy-indicative terms and found that a language model of documents containing "controversy" keyword directly is as helpful for this problem as complicated Wikipedia-based controversy features and more effective than existing lexicons. We finally compare the two models, kNN-WC and CLM, which have a few complementary properties to each other. kNN-WC model

has an advantage of being able to retrieve specific topics as a reference with the risk that contention signals of many specific topics could be missing. Regarding that, we have addressed a technique to alleviate this issue via smoothing. CLM is more efficient to compute, and does not suffer from the sparse "contention" signals as they examine the language of the document. While kNN-WC is tuned to capture the mention of controversial topics in the document, CLM considers the balance between the controversial and non-controversial language of the document.

Table 5.10: ClueWeb document "clueweb09-en0005-61-08920" was correctly labeled as controversial by kNN-WC while CLM labeled it as non-controversial. The above table indicates the document text (after removing the html tags and the boilerplate) whose controversial terms are annotated by CLM with color meaning: controversial > somewhat controversial. The table on the bottom shows the top 20 retrieved Wikipages by TileQuery method along with M and C score.

In 1986, government scientists wrote a letter to the British journal Lancet and acknowledged that abortion is a cause of breast cancer., They wrote, "Induced abortion before first term pregnancy increases the risk of breast cancer.",(Lancet, 2/22/86, p. 436) As of 2006, eight medical organizations recognize that abortion raises a woman's risks for breast cancer, independently of the risk of delaying the birth of a first child (a secondary effect that all experts already acknowledge).

An additional medical organization, the Association of American Physicians an Surgeons, issued a statement in 2003 calling on doctors to inform patients about a "highly plausible" relationship between abortion and breast cancer., General counsel for that medical group wrote an article for its journal warning doctors that three women (two Americans, one Australian) successfully sued their abortion providers for neglecting to disclose the risks of breast cancer and emotional harm, although none of the women had developed the disease. Click here for more

---

Rank	Wikipage Title	M score	C score
1	Abortion-breast_cancer_hypothesis	1789961	0.000
2	Risk_factors_for_breast_cancer	0	0.002
3	Breast_cancer	12529	0.012
4	Breast_cancer_awareness	0	0.001
5	Joel_Brind	0	0.001
6	Voice_for_Life	0	0.001
7	Crisis_pregnancy_center	0	0.030
8	Sharsheret_(organization)	0	0.000
9	Cancer	5469	0.020
10	Abortion	3743570	0.296
11	Susan_G._Komen_for_the_Cure	32	0.003
12	Breast_cancer_research_stamp	0	0.000
13	Alcohol_and_breast_cancer	0	0.000
14	Triple-negative_breast_cancer	0	0.000
15	Dressed_to_Kill_(book)	0	0.001

Table 5.11: ClueWeb document "clueweb09-en0007-98-30872" was correctly labeled as controversial by CLM while kNN-WC labeled it as non-controversial. The above table indicates the document text (after removing the html tags and the boilerplate) whose controversial terms are annotated by CLM with color meaning: controversial > somewhat controversial. The table on the bottom shows the top 20 retrieved Wikipages by TileQuery method along with M and C score.

... Mission statement free homeopathy educational materials.
This is an open homeopathy project for all by all. Let s make all aware of the wonders of homeopathy. Do it yourself approach for healthy and holistic living.
Homeopathy restore health rapidly gently and permanently. Homeopathy medicines are patent free inexpensive and harmless.

First aid situations or acute illnesses treat yourself by homeopathy
classical homeopathy approaches as well all unconventional approaches are equally respected and welcome here please feel free to contribute and share your knowledge and experience picture of this moment.
This site provides only educational materials all advices given here are only for educational purpose.

Rank	Retrieved Wikipage	M Score	C Score
1	Waldorf_education	196630	0.091
2	List_of_alternative_therapies_for_developmental_and_learning_disabilities	0	0.000
3	Edward_hamilton_(homeopath)	0	0.000
4	Tadepalle,_krishna	0	0.001
5	Nelsons_(homeopathy)	0	0.001
6	Educational_research	0	0.001
7	Arthur_lutze	0	0.000
8	Faculty_of_homeopathy	0	0.001
9	The_forbidden_education	0	0.000
10	Efterskole	0	0.000
11	Gheorghe_jurj	0	0.000
12	Puget_sound_community_school	0	0.000
13	George_vithoulkas	0	0.004
14	Universidad_del_sagrado_corazon	0	0.000
15	Rajesh_shah	0	0.000
16	Glossary_of_alternative_medicine	0	0.001
17	Beykent_educational_institutions	0	0.004
18	Motiwala_education_and_welfare_trust	0	0.004
19	Educational_psychologist	0	0.002
20	Mel_wasserman	0	0.000

CHAPTER 6

ESTIMATING TEMPORAL CONTROVERSY TRENDS

6.1 Introduction

6.1.1 The Dynamic Nature of Controversy

Naturally, the level of controversy changes as the topic evolves over time and the discourse of the topic develops. People's attention and interest in the matter change over time as well, which naturally affect the amount of online discussions on the topic. The topic could get more heated as it goes more "viral" or it can naturally die over time simply because there is no further development or because people simply become bored of it.

In a case study of controversial events, Cramer (2011) found that terms that describe the *Busang case* (Depalma, 1997) have shifted from "dispute" and "controversy" to "saga" and "scandal" over time. This demonstrates how the nature of a controversy changes as it develops. This phenomenon is demonstrated by our study that presented a plot of the daily level of controversy measured in Twitter in Figure 6.1 (Jang et al., 2017). It shows that some controversies are more ephemeral than others. For example, "The Dress" controversy, the controversial photo that went viral when people disagreed on its colors on Twitter, was no longer controversial on Twitter after only a few days as most people stopped caring. On the other hand, "2016 U.S. Presidential Election" had a longer span of controversy, a longer-lasting effect than "the Dress."

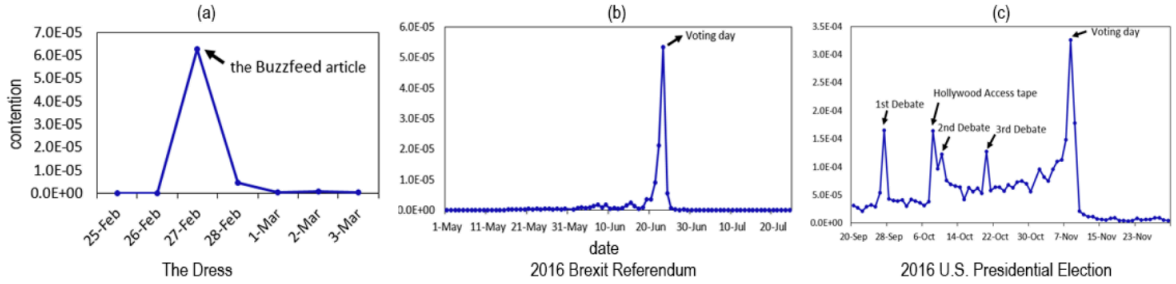

Figure 6.1: Controversy computed by P score (Jang et al., 2017) among all daily tweets by date for The Dress (left), Brexit (center) and 2016 US Elections (right), reported among those Gardenhose tweets with an explicit stance. Notable peaks are annotated with associated events around that time. All dates are in UTC (in 2016).

6.1.2 True Controversy beyond Observed Conflicts

Online controversies often drive digital attention. Therefore, the level of people's attention at that time is an important factor that contributes to the amount of presentation of controversial discussion online. Because people tend to have a limited amount of attention, a newer controversy constantly fights for people's attention on the Web. For such reason, when a certain controversy is not surfaced online at the moment, it does not always mean that the topic is no longer controversial. It might just mean that the controversy is currently latent, relatively out of public interest.

The existing topic controversy models focus on analyzing present signals and do not consider this phenomenon, hence do not see beyond the observed conflicts. When we don't observe controversial signals from the given platform at a given time, does that mean that the topic is not controversial or is "latently" controversial such that we just don't see it at that time? For example, In Figure 6.1, we could reasonably assume that 'The Dress" was probably no longer controversial after a short time whereas "2016 Brexit Referendum" and "2016 US Presidential Election" were more controversial for a while.

In retrospect, this issue was similarly observed in the kNN-WC algorithm (Chapter 4), when the automated controversy scores such as M scores and C scores are underrated for Wikipages that receive less attention and that have similar topics to the page where the editors have disputed the issues. Existing approaches have been more focused on analyzing the controversial signals that are currently available and do not differentiate these cases to predict the true controversy level looking beyond the observed conflicts.

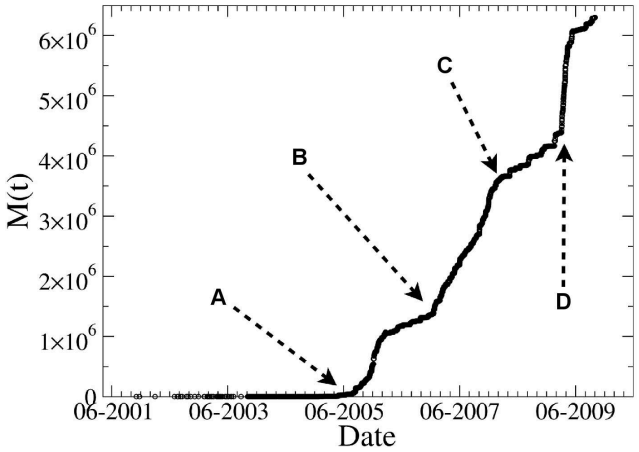

Figure 6.2: "Time evolution of the controversy measure of the article about Michael Jackson. A: Jackson is acquitted on all counts after five month trial. B: Jackson makes his first public appearance since the trial to accept eight records from the Guinness World Records in London, including Most Successful Entertainer of All Time. C: Jackson issues Thriller 25. D: Jackson dies in Los Angeles." Source: http://wwm.phy.bme.hu/

6.1.3 Monotonicity of Controversy Scores in Wikipedia

Because *time* was not directly modeled in the existing approaches, they often have a monotonic property over time. For example, M score (Yasseri et al., 2012), one of the successful methods that estimates the level of controversy in Wikipedia in proportion to the number of mutual reverts among credible editors, uses the edit history accumulated over time. Hence, the longer the edit history is, the more likely we are to have mutual reverts, and the more likely the M score is to get bigger. This

is demonstrated in Figure 6.2 via a topic that was once highly controversial: Michael Jackson. Figure 6.2 shows the evolution of the M score on "Michael Jackson". The graph shows that the controversy score has monotonically increased every time there is a new controversial event added on to the article up until the point "D" where he died. However, ever since then the controversy score still remains as high as D (or higher) until later in 2012.

Some approaches are not monotonic as their scores are normalized by the number of editors who contributed to the page, which increases over time. Dori-Hacohen argued that P score (2017) can go up and down as time goes by, because they focus on the ratio of editors who are in conflict compared to the entire editor population on the topic. Their intuition is that over time if they have more editors who are not involved with disputes, the controversy score will be decreased because a lower ratio of people engage in the disputes. However, this requires more people to actively engage in non-contentious activities to cancel out the level of controversy. If simply no one cares to talk about the topic anymore, it still remains controversial over time.

6.2 A Case Study of Time-window-based M Score

As the monotonicity of M score was due to the fact that we consider all the edit-history that has accumulated to the date, a straightforward solution to this issue is to consider only a given window of time to estimate the controversy for that time. We downloaded a Wikipedia dump of 2018-06-01 to generate a M score trend over the past 18 years since the existence of Wikipedia. We analyzed the top 100 most controversial topics by the accumulated M score. It turns out that a time-window-based M score has the opposite problem: while the monotonically-increasing M scores that were computed from the all history tend to be overrated, this version of M score seems to be largely underrated. The M score trend for most topics shows a burstineses

where there are a few spikes in the trend line while having zero points most of the time.

While the controversy trend line is known to be bursty both in Wikipedia and social media, we learn that the burstiness comes from different reasons based on the nature of their platform. As social media is a place where users can post any opinion any time they want, the similar arguments and opinions can take place and be reproduced over and over as much as users would like to speak out. Usually, on social media we observe users' opinions posted on the controversial topic as part of the *reactions* to a certain event that happened during that time. Most events are temporal, which create bursty trend lines as shown in Figure 6.1. On the other hand, in Wikipedia, the dispute signals are not from personal reactions but rather from arguments that occur as part of the collective effort towards generating unbiased content on that topic. Due to this nature, most disputes of the topic usually occur upon document creation, or controversy creation. Once the Wikipage is matured, the article is maintained with fewer disputes, showing only a few or none for most of the time unless a new controversial event occurs. Even then, the fundamental discussion on the controversial topic has already been settled, the score in the later year is rarely not even remotely close to the peak at an early year (refer to "Elvis Presley" (top right) and "Falun Gong" (bottom left) in Figure 6.3).

We argue that in order to correctly estimate the controversy value at a given time, we need to consider the signals observed within a window of time as well as the overall history of the controversy. In this work, we assume that the dispute signals we observe through online activities are only observed and biased samples of all controversial disputes in the real world.

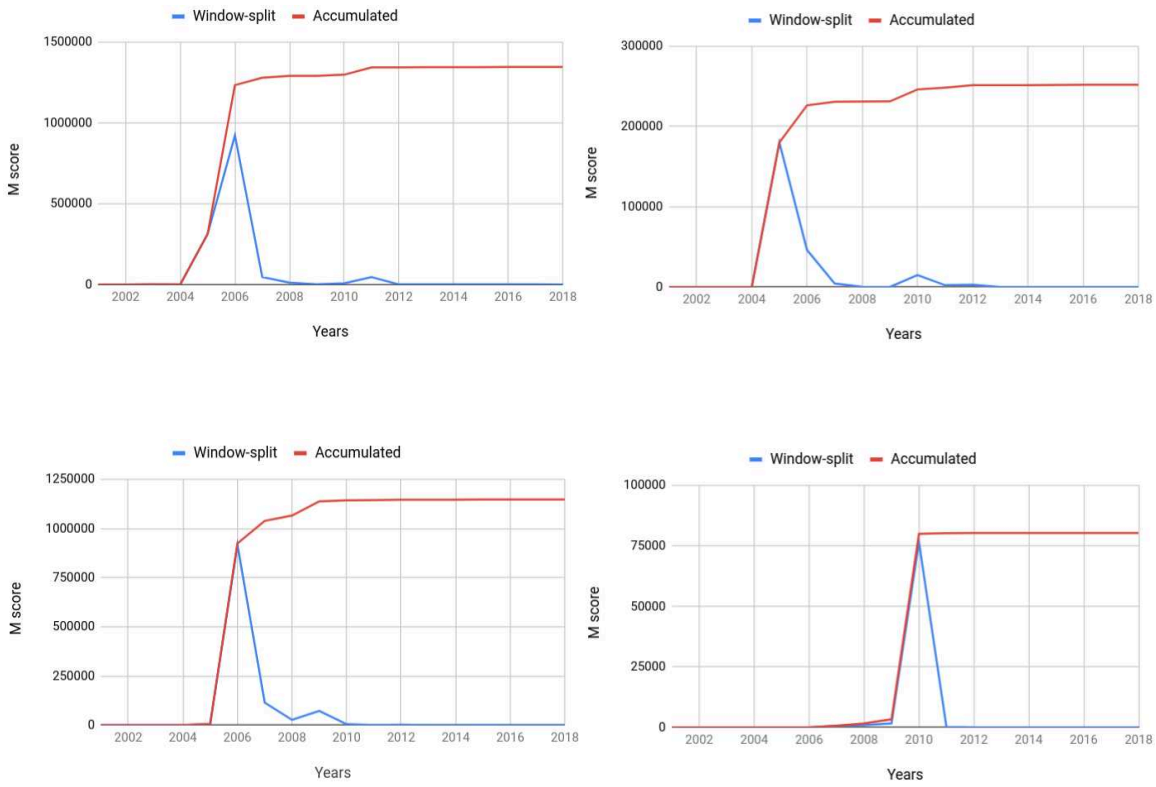

Figure 6.3: The time-window-based M score with window of 1 year (blue line) and its cumulative trend (red line). The top left (Abortion), the top right (Elvis Presley), the bottom left (Falun Gong), the bottom right (2010 Fifa World Cup).

6.3 Estimating True Controversy

In order to estimate the *true* controversy score at time y (as in year) from the observed disputes, we hypothesize that there are two factors that determine the true controversy score: contention and public interest. In our previous work that is not part of this thesis, we have shown that controversy should be modeled within a population and proposed a model of controversy should comprise at least two primary dimensions, the level of contention and importance of the topic within a given population (Jang et al., 2017). In the previous work, "importance" was conceptually defined and estimated via the number of people who discuss the topic. Similarly, we define the controversy of a topic at any time t to be modeled as two factors, contention and

public interest. Because we are interested in a general-purpose controversy function, we assume that the given population is a general, all encompassing population.

Finally, we model the probability of controversy with a given topic T and a given time Y. Let C a binary random variable, which denote the presence of controversy. Similarly, let $Cont$ and I be binary random variables, which denote the presence of contention and public interest of topic T. We model $P(Cont|\theta)$, where $\theta = \{T, Y\}$ as the probability that topic T is controversial within the population Y. Our model hypothesizes that the probability of controversy given T and Y is the joint probability of two dimensions: contention ($Cont$) and public interest (I):

$$P(C|\theta) = P(Cont, I|\theta)$$

Here, $P(Cont, I|\theta)$ can be further decomposed as following:

$$\begin{aligned} P(Cont, I|\theta) = \frac{P(Cont, I, \theta)}{P(\theta)} = \frac{P(I|Cont, \theta) \cdot P(Cont|\theta) \cdot P(\theta)}{P(\theta)} \\ = P(I|Cont, \theta) \cdot P(Cont|\theta) \end{aligned} \tag{6.1}$$

To compute $P(I|Cont, \theta)$, the correlation between contention and public interest has to be identified. While it is difficult to estimate the exact correlation in the real world, we assume that contention and public interest are independent of each other, consisting of orthogonal dimensions of controversy. We therefore let $P(I|Cont, \theta) = P(I|\theta)$.

$$P(C|T, y) = P(Cont|T, y) \cdot P(I|T, y) \propto \mathcal{C}_y = c_y \cdot p_y \tag{6.2}$$

where $\mathcal{C}_y$ is the score that indicates the level of true controversy at a given time y, c_y is the true level of contention, and p_y is the true level of public interest.

We note that the existing controversy scores that are analyzed from dispute signals are not the true controversy score $\mathcal{C}_y$, but the *observed* controversy score $\hat{\mathcal{C}}_y$ and clearly distinguish the two scores: $\mathcal{C}_y \neq \hat{\mathcal{C}}_y$.

In the following section, we introduce models to estimate the true controversy score from the observed controversy score and the level of public interest.

6.4 Methods

6.4.1 Models for true contention from observed controversy

$$\hat{\mathcal{C}}_y = \hat{c}_y \cdot \hat{p}_y \tag{6.3}$$

where $\hat{\mathcal{C}}_y$ is the observed controversy score of a given topic at time y, $\hat{c}_y$ is the observed level of contention, $\hat{p}_y$ is the observed level of public interest. Wikipedia controversy scores have an especially severe gap between the observed controversy level and the true controversy level because once the dispute has been settled, the same dispute are not likely to be duplicated. In the meantime, public interest, which is temporal reactions to the topic, does not have such constraint. Hence, we assume that the observed level of public interest is relatively reliable and set $\hat{p}_y = p_y$. So,

$$\hat{\mathcal{C}}_y = \hat{c}_y \cdot p_y \tag{6.4}$$

Max Contention - interest (MCI) Model: In this model, we assume that the true latent contention at a given time is the same as the maximum level of observed contention. This assumes that the topic that was once highly contentious remains latently that contentious. This approach assumes that the topic always has a potential to be as contentions as it has historically been while high interest on the topic could activate the controversy with the latent contention.

$$c_y = \max_{i=1..y} \hat{c}_i = \max_{i=1..y} \frac{\hat{\mathcal{C}}_i}{p_i} \tag{6.5}$$

The final true controversy by MCI is obtained by the following:

$$\mathcal{C}_y = \frac{\hat{\mathcal{C}}_j}{p_j} \cdot p_y \tag{6.6}$$

here j is a time when M score was at its maximum and defined as:

$$j = \underset{x \in \{1..y\}}{\operatorname{argmax}} \frac{\hat{\mathcal{C}}_x}{p_x}$$

Accumulated Contention - interest (ACI) Model: In this model, we assume that the true contention is the same as the accumulated level of observed contention. The difference between this model and accumulated M scores (Section 6.1.3) is that in this model, only the level of contention is accumulated whereas the level of public interest is also accumulated in the latter. Therefore, while accumulated M score has a monotonically-increasing trend line, the trend from this model is not monotonically-increasing as the level of public interest fluctuates. The true controversy is obtained as follows:

$$c_y = \sum_{i=1}^{y} \hat{c}_i = \sum_{i=1}^{y} \frac{\hat{\mathcal{C}}_i}{p_i} \tag{6.7}$$

However, public interest may not perfectly align with the observed controversy from Wikipedia because usually there is some delay before the controversy is observed in Wikipedia. Such delay could particularly be detrimental in this method where the true contention is computed point-wise on a daily basis and many points will have low observed controversy scores, most of which are themselves unreliable. Hence, instead

of using public interest on the same day, we use the average value of public interest accumulated until that day as type of a smoothing.

$$\overline{p_y} \approx \underset{i=1..y}{avg}\, p_i \tag{6.8}$$

The final true controversy by ACI is obtained by the following:

$$\mathcal{C}_y = \sum_{i=1}^{y} \frac{\hat{\mathcal{C}}_i}{\overline{p_y}} \cdot p_y = \frac{\sum_{i=1}^{y} \hat{\mathcal{C}}_i}{\overline{p_y}} \cdot p_y \tag{6.9}$$

Window Contention - interest (WCI) Model: In this model, we assume that the true latent contention constantly changes over time and can be estimated from looking at a window of history of the observed contention.

$$c_y = \underset{i=y-w..y}{avg}\, \hat{c}_y = \underset{i=y-w..y}{avg} \frac{\hat{\mathcal{C}}_i}{p_i} \tag{6.10}$$

$$\mathcal{C}_y = \underset{i=y-w..y}{avg} \frac{\hat{\mathcal{C}}_i}{p_i} \cdot p_y \tag{6.11}$$

6.4.2 Obtaining Observed Controversy

For the observed controversy $\hat{\mathcal{C}}$, we use M score. M score takes into the number of disputes that have occurred and has both *contention* and *interest* entangled in their score while it considers the number of the editors and the minimum reputation score of editors for each mutual revert. While the level of contention is proportional to the number of mutual reverts, the level of public interest is proportional to the number of editors.

6.4.3 Obtaining Public Interest

To estimate the level of public interest on the topic, we resort to Google Trends service[1]. Google Trends is a website that analyzes and shows the popularity of the search queries in Google Search. The website allows a comparison of the search volume of two or more queries over time. We adopt the trend line provided by Google Trends as a reasonable estimation of public interest on the topic. Originally, Google Trends only provides a relative trend line that is normalized by the maximum volume point during the time period within a given topic, or the multiple topics of interest. Hence, this does not give us absolute values that are comparable across multiple topics (Figure 6.4). Therefore, to obtain the trend line values that are comparable across all topics, we convert the trend lines into the same scale based on the fact that comparisons of two trends are transitive. We turn this into a problem of generating one connected graph with all nodes where each topic of interest corresponds to a node and two nodes are connected if the comparison trend lines between the two topics is obtained. Once all topics are connected via a comparison trend line, we convert the trend lines of all topics into the points in the same comparable space.

6.5 Model Validation: A Case Study

We validate our time controversy models via a qualitative analysis. Evaluating the controversy trend over the last 14 years is tricky. While the previous controversy dataset relied on human judgment to identify whether a topic is controversial, it would be difficult to find reliable annotators that can correctly recall the level of controversy of the given topic for the past 14 years. Hence, we resort to examining various cases to validate our model.

[1] `https://trends.google.com/trends`

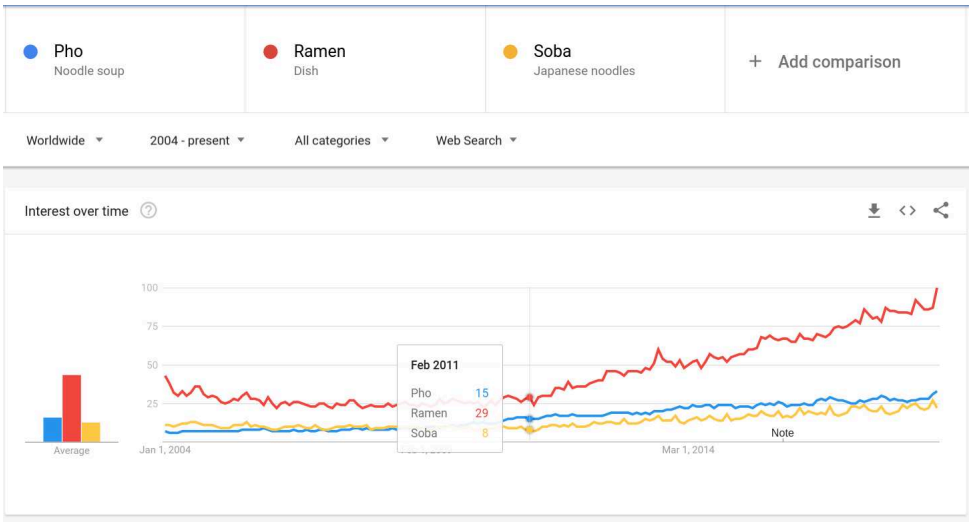

Figure 6.4: A screenshot of Google Trends that shows a trend line comparison among three queries, Pho, Ramen, and Soba. While the trend line shows the relative comparison among the queries, the absoulte value of each trend line is unknown.

6.5.1 Abortion

"Abortion" is a well-known controversial topic. In Wikipedia, the most disputes have been occurred in 2005 and 2006 showing a high peak during those early years. Since 2007, the level of controversy significantly dropped until 2012 when there is no controversial signal anymore. This is one of the common pattern shown for many long-term controversial topics. In the mean time, public interest started very high in the early years and has also decreased over time with some fluctuation. Figure 6.5 shows the predicted true controversy trend line using AIC, MCI, and WIC, respectively. While both ACI and MCI constantly predicted "Abortion" to be highly controversial at all times, WCI predicted that the topic is no longer controversial after 2012 as the topic did not show any contention in the 5-year-window. As a long-time ethical controversy, there is no clear evidence or reason that suggests that the level of controversy has increased over the last 14 years as ACI suggests nor that it is no longer controversial as WCI suggest in 2018. Hence, the trend by MCI reasonably

suggests that "Abortion" is still highly controversial with small fluctuations along with public interest.

6.5.2 Kim Jong-il

Kim Jong-il was the second Supreme Leader of North Korea, who served since the death of his father Kim Il-Sung and until his own death in 2011. Kim Jong-il had been involved with many controversial issues and accusations of human rights violation such as mass starvation, executions, and forced labor (Wikipedia, 2019b).

This Wikipage was created in 2002, and started getting serious editors' contributions from 2003. This topic's M score also follows the same pattern as "Abortion" where controversial disputes have occurred while this topic was actively being curated in the early few years. The mutually-reverted edits suggest the controversy between editors included whether he "ruled" or "led" the country and the discussion over Kim Jong-il's intention with regard to North Korea's relation to South Korea. When he died in 2011, public interest spiked.

Figure 6.8 shows the predicted controversy trend from MCI and WCI with a window of 5 years. We omit the trend from ACI as it showed the same pattern as MCI because the maximum contention was close to the accumulated level of contention. While the accumulated M score suggests that Kim Jong-il is still controversial in 2018 as it would for any topic that was once controversial, and the window-based M score suggests that Kim Jong-il is not controversial even in 2011 when he died, and the trend from MCI suggests that Kim Jong-il is still somewhat controversial while a gradually decreasing pattern after being particularly controversial in the year he died. The trend from WCI shows that Kim Jong-il was controversial over the years while he was alive, but no longer controversial since he died. Kim Jong-il is still a somewhat controversial topic in 2018 as his policies and remarks are still being quoted when

Figure 6.5: The trend of **Abortion** from AIC, MCI, and WCI with a window of 5 from the top. The blue trend line indicates the predicted controversy trend line with AIC. The red bars indicate the M score in the given year. The grey line shows public interest from Google Trends.

his son, “Kim Jong en”, who is another controversial topic himself is being discussed (Denyer, 2018).

6.5.3 Taiwan

Taiwan was one of the top 50 controversial topics in Wikipedia by M scores. While many controversial topics have a pattern of having high controversy scores in the early years upon the document creation and not having further signs because the topic has been saturated (e.g., Abortion and Kim Jong-il), this topic showed relatively consistent level of controversy over the 14 years. The mutually-reverted edits such as “Chinese people <-> Taiwanese people”, “Mainland China <-> Mainland China and Taiwan”, suggest that the main controversy around this topic has been whether or not to view Taiwan part of China.

6.5.4 Race and Intelligence

The link between race and intelligence is a highly controversial debate since at least the invention of the intelligence test. The controversy includes whether and to what extent genetic factors and environmental factors affect in the intelligence test scores as well as the definitions of what “race” and “intelligence” are. The mutually-reverted text mainly includes argument on the inclusion and deletion of uncredible sources of the claims that could bias the readers’ judgment on the issue. In Wikipedia, the topic was shown to be highly controversial for the first 5 years upon document creation, and the observed controversy trend has waned since then. This is one of the most common patterns that we see in M scores.

In this topic, the trend lines by ACI, MCI, and WCI, respectively suggest different trends. ACI suggests that the true controversy consistently increases over time. MCI suggests that the trend has been fluctuating while peaking together with the peaks of public interest, while remaining at a consistent level of controversy over time. WCI with window of 5 years suggests that the trend slowly decreases over time. This

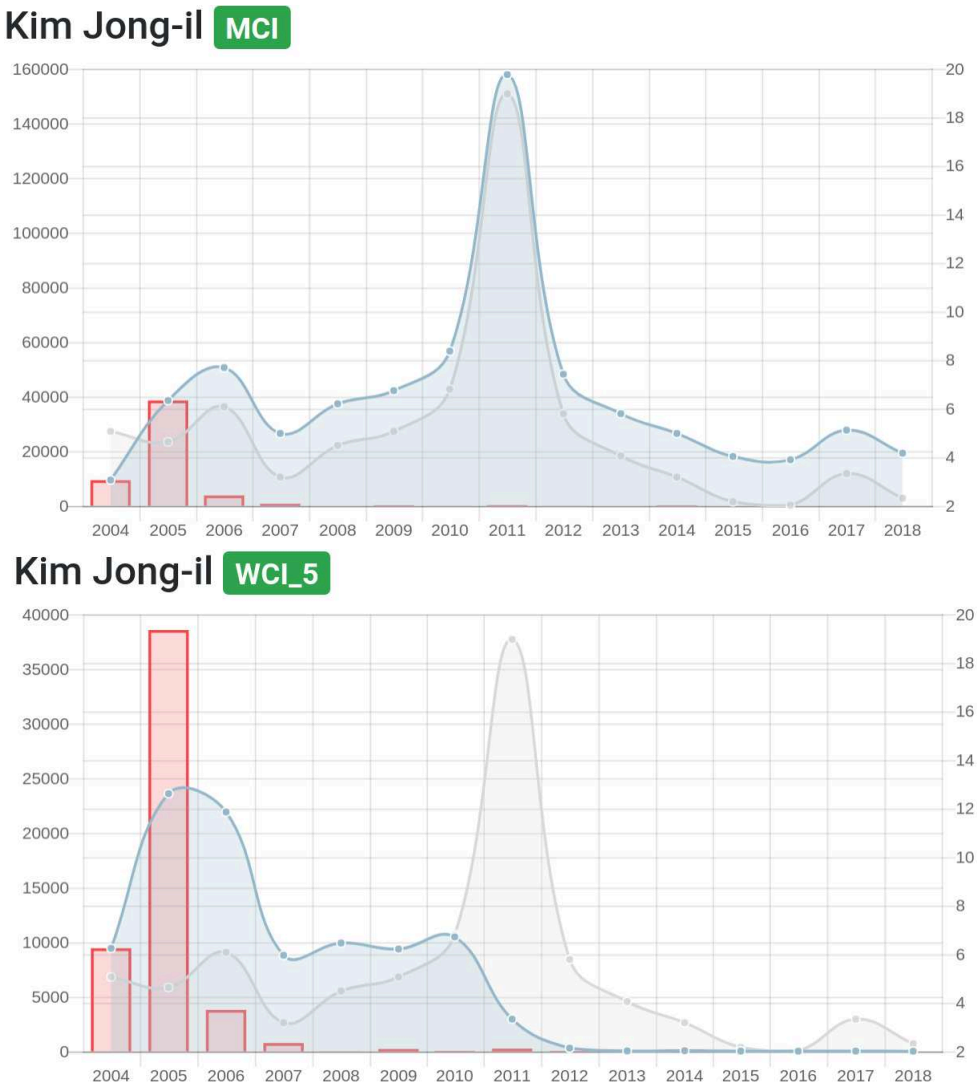

Figure 6.6: The trend of **Kim Jong-il** from MCI and WCI with a window of 5 from the top. The blue trend line indicates the predicted controversy trend line with AIC. The red bars indicate the M score in the given year. The grey line shows public interest from Google Trends.

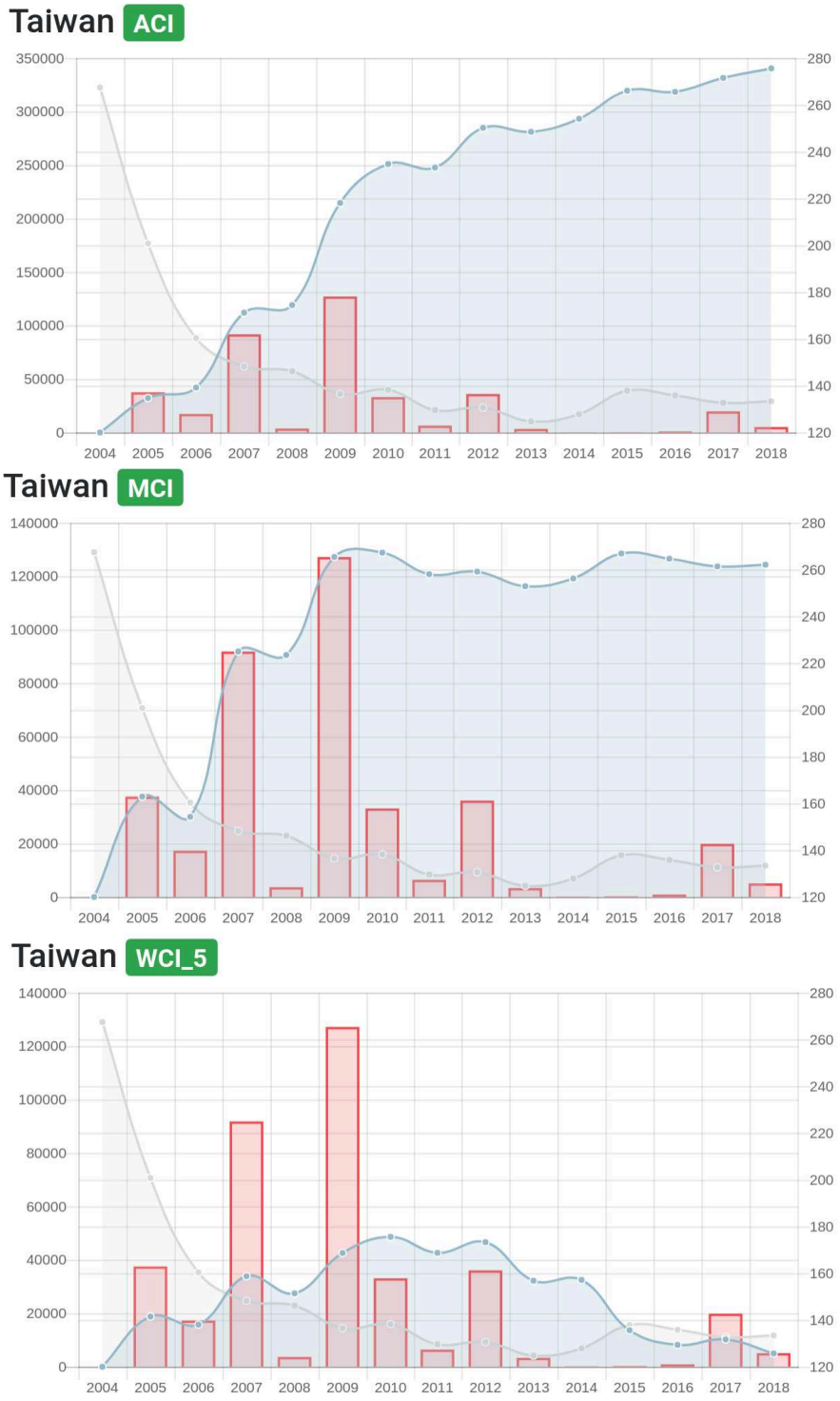

Figure 6.7: The trend of **Taiwan** from MCI and WCI with the window of 5 from the top. The blue trend line indicates the predicted controversy trend line with AIC. The red bars indicate the M score in the given year. The grey line shows public interest from Google Trends.

controversy seems to have remained controversial until recently. MCI and ACI both suggest that the controversy peaked in the following four years: 2007, 2009, 2013, and 2017. We examine if there is a controversial event that can explain why this topic was particularly controversial in each year.

- In 2007, James Watson, a Nobel-prize winning scientist stated in an interview that research has suggested without any scientific evidence that for genetic reasons Africans have lower intelligence than Europeans. He was forced to retire from Cold Spring Harbor Laboratories after his statement.

- In 2009, *Science's Last Taboo* was a British TV show about race and intelligence broadcast on Channel 4 in 2009. This TV show caused controversy from statements claiming that Africans are less intelligent than Caucasions and East-Asians.

- In 2017, Rindermann et al., (2016) published a new study that attempted to replicate the earlier findings of Snyderman & Rothman (1988) by surveying 71 psychology experts and claiming that education is the most important factor of the intelligence score gaps among the races followed by genetics. This study sparked several controversial discussion thread in Reddit (Reddit, 2018a,b,c).

6.6 Conclusion

In this chapter, we argue that the controversy scores that existing models generate by analyzing dispute signals reflect the level of *observed* controversy and they do not accurately reflect the *true* controversy score in real life. We distinguish the two concepts and propose to estimate the true controversy scores that change over time from the observed controversy scores. We propose a model that considers "contention" and "public interest". We first obtain the observed contention scores by separating

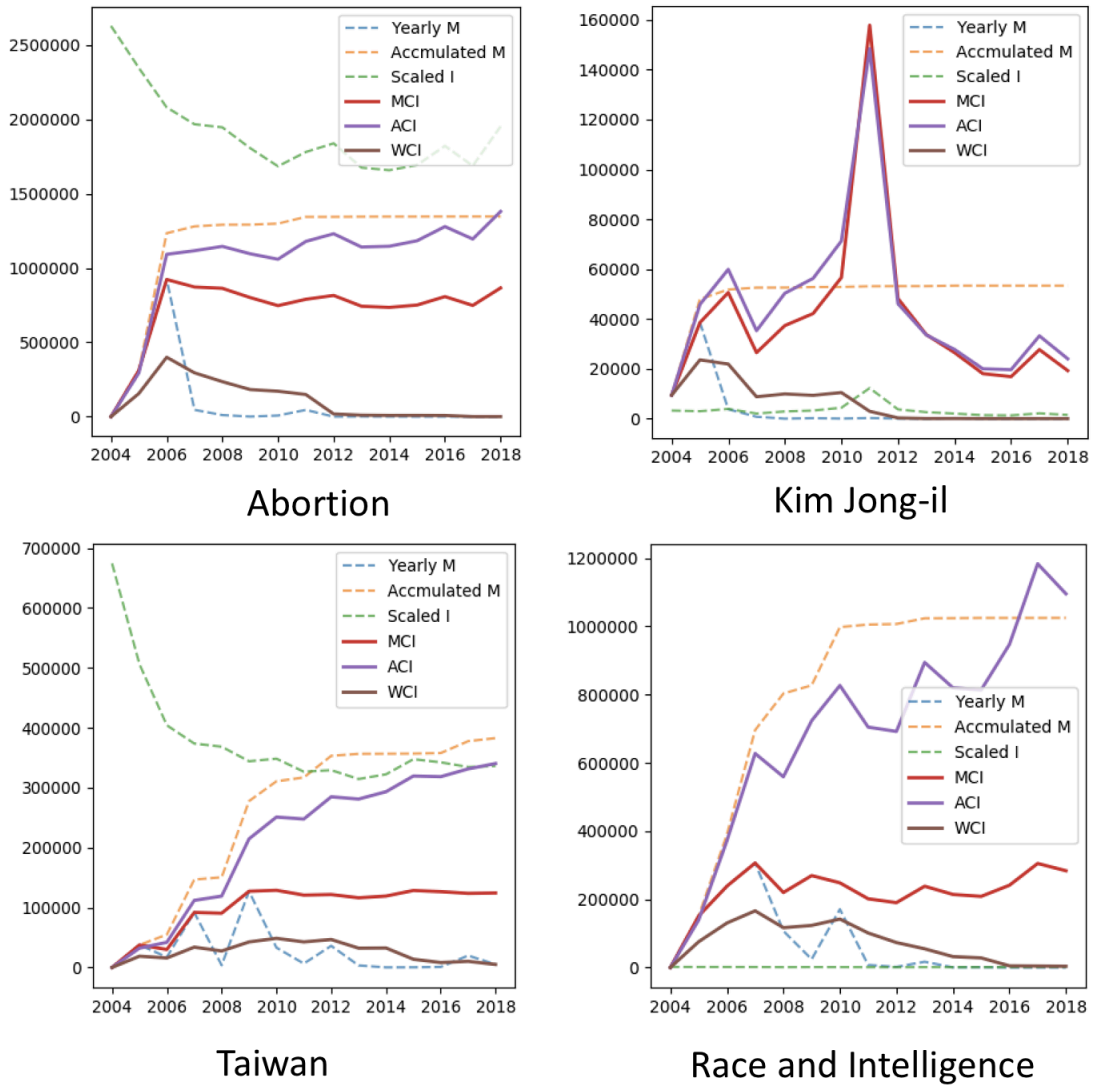

Figure 6.8: The trend of a yearly M score, accumulated M score, public interest, MCI, ACI and WCI. The raw score of public interest was very low compared to the other scores, we scaled it up by multiplying the tenth of the average of the public trend.

the component of popularity from M scores. We then introduce three methods – MCI, ACI, and WCI – that multiply the true contention with public interest. Each method estimates true contention from observed contention differently by taking the maximum contention, the accumulated contention, and the average contention in a moving window. We validate our methods via a case study. We find that many long-term controversial topics share a tendency that the observed controversy scores are high upon the Wikipedia article creation until the topic becomes more mature and that fewer edits are made. Due to this reason, while WCI is more adaptive and suitable to predict the controversy trend more accurately for short-term controversial topics, WCI seems to underrate the controversy scores as the moving window no longer includes this early period for long-term controversial topics. ACI and MCI show similar patterns for the topics that have few dominant peaks where the maximum contention and the accumulated contention is almost the same. While MCI and ACI generate a similarly fluctuating pattern, they differ in the pattern of the overall trend over time. ACI generates trends that controversy increases over time often even with a reduced amount of public interest in the later time; MCI generates a relatively consistent trend. Without any evidence or reason to believe that the controversy necessarily have increased in the topics examined, we find that MCI generates the most reasonable trend that reflects the true controversy.

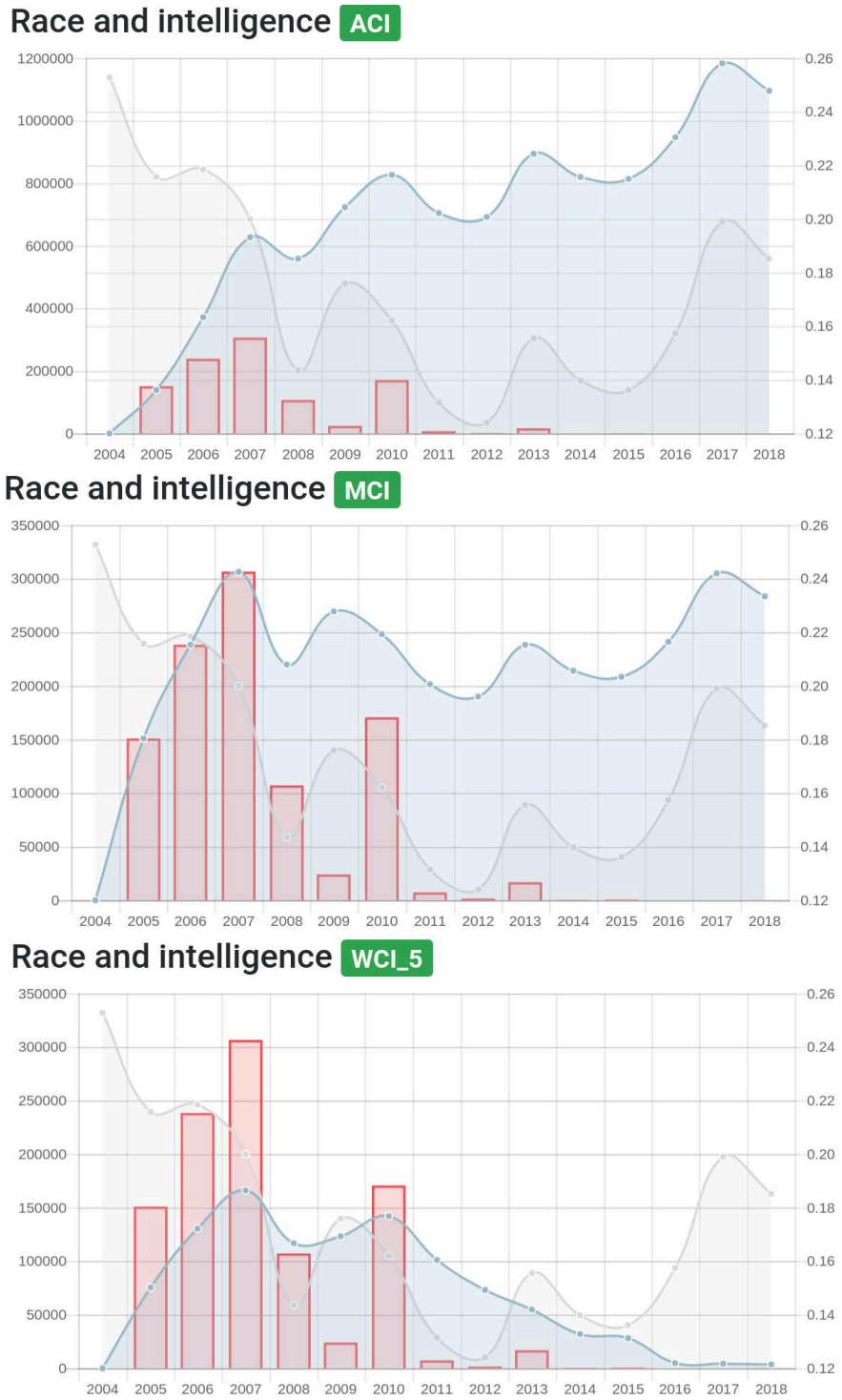

Figure 6.9: The trend of **Race and Intelligence** from ACI, MCI and WCI with the window of 5 from the top. The blue trend line indicates the predicted controversy trend line with AIC. The red bars indicate the M score in the given year. The grey line shows public interest from Google Trends.

CHAPTER 7

EXPLAINING CONTROVERSY ON SOCIAL MEDIA

7.1 Introduction

Online controversies often emerge and evolve quickly due to the nature of social media. These platforms force users to be concise and allow them to be casual, requiring less effort to post something on Twitter than other sources, such as Wikipedia or blogs. While existing techniques enable us to identify *whether* a topic is controversial, understanding *why* it is controversial is still left as work for users. For instance, consider a following scenario: A person discovers a new hashtag movement `#TakeaKnee`[1] on Twitter but does not know what it is about or why it is controversial at all. How would she search for people's opinions to better understand the conflicting stances on this topic?

One straightforward approach to this problem would be for the user to search the topic and manually scan the search results until she has read enough conflicting tweets to understand the controversy. However, current search systems make this navigation difficult due to the filter bubble effect (Ingram, 2016). For example, the top posts are likely to be the ones that the user agrees with because her friends liked the posts or because she or her friends follow the authors.

Another strategy for navigating Twitter is to identify a few key hashtags that indicate stances and then search for posts that contain them. As people are forced to write posts under the strict character limit, certain hashtags are utilized as self-created labels for their opinions (e.g., `#imwithher` in support of Hillary Clinton

[1]This was prevalent during the US national anthem protests that began in 2017.

or #MAGA in support of Donald Trump during the 2016 US presidential election). However, because the use of hashtags (even the ones that have seemingly contain obvious stances) are known to be noisy (Mohammad et al., 2016b), the user must still carefully read through each tweet. More importantly, she has to go through a large number of noisy tweets that are not useful to understand the controversy while using her own judgment to identify their stance (if they even have one). This process requires substantial effort, critical reasoning, and phenomenal patience. It is clear that users could benefit from automating this process.

We propose a technique that generates a stance-aware summary by selecting the top tweets that best explains a given controversy.

7.2 Related Work

As having at least conflicting two stances is a major characteristics that defines controversy (Jang et al., 2017), we generate a stance summarization on social media to explain why the given topic is controversial (Chatper 7). We survey the related work in this area.

7.2.1 Stance Detection on Twitter

Stance classification on Twitter has two main tasks: (1) classifying the text's stance (against, favor, or neutral) given a topic, and (2) classifying the twitter users' stances. The former task drew attention when 2016-SemEval Task 6 released a dataset of tweets with stance annotations (Mohammad et al., 2016b). The results of various approaches were shared after the competition (Mohammad et al., 2016c), and later more successful approaches were proposed including one that uses a bi-directional conditional LSTM for classifying the stance and opinion target on Twitter (Augenstein et al., 2016). For the latter type of task, Johnson and Goldwasser developed a method to classify stances of politicians on Twitter using relational representation (Johnson

and Goldwasser, 2016). While stance detection is closely related to our problem, our goal is not to accurately classify the stances of all tweets. Our problem is also more robust to misclassification errors of stances as we take the tweets with highest stance confidence as part of the summary.

7.2.2 Twitter Summarization

There has been much work on summarizing Twitter postings through most of them focuses on summarizing events (Sharifi et al., 2010; Duan et al., 2012; Chakrabarti and Punera, 2011; Inouye and Kalita, 2011; Yulianti et al., 2016). Inouye et al. 2011 compare multiple summarization algorithms for Tweet data, and their extensive experiments suggest that the SumBasic algorithm (Nenkova and Vanderwende, 2005) produced the best F1-result in human evaluation. SumBasic is a summarization algorithm that uses the term frequency exclusively to create summaries. As a simple system based on word frequency in the document set, SumBasic outperformed any other complex system at the time. SumBasic computes the best k posts from the input documents that contain a lot of high frequency terms. We choose SumBasic as our baseline method.

Some work has focused on generating contrastive summaries from opinionated text (Paul et al., 2010; Guo et al., 2015). Particularly, Guo et al. studied tweet data to find a controversy summary. They find a pair of contrastive opinions by integrating manually-curated expert opinions and clustering the pairs to generate a summary. However, their model needs curated expert opinions, which requires constant human effort to maintain as the topic evolves.

Table 7.1: An example of good (top) and bad (bottom) summary tweets on "Abortion" posted on Nov 4, 2016. The good summaries are selected from our method. Examples of stance hashtags are marked in bold.

• We know it's not okay that for 40 yrs politicians have denied a woman coverage of abortion just because she's poor #**BoldTheVote** #**BeBoldEndHyde**
• Read the whole story about #HarvardSoccer before forming idiotic tweets. Don't support #**RapeCulture** by calling it #LockerroomTalk
• Hillary Clinton voted no to banning late-term abortions, even though over 80% of Americans support the ban. #**VoteProlife**
• lmaoaoao b**** i would did the abortion myself right there lmaoaoao
• before I formed you in the womb I knew you jer 1:5#**prolife** #**Defundpp** [URL] #**UnbornLivesMatter**
• Abortions: the new fall trend in religious circles [URL]
• Could you imagine crying over ur uni stopping anti abortion protests, if you're so pro life then go and f***ing get one?

7.3 Approach

7.3.1 What Makes a Good Summary Tweet?

In order to design a ranking model that ranks the tweets by how likely a tweet is to be part of a good summary, we first need to discuss the definition of a "good summmary" for controversy.

One of the primary aspects for the definition of controversy has been "contention". This suggests that in order to understand controversy, one needs to understand what causes disputes or conflicts between the two parties. Based on that, we define a good controversy summary as a description that effectively captures the representative arguments of two communities that take conflicting stances with each other. To obtain an intution on the characteristics of a good summary, we manually examined many examples on Twitter on controversial topics.

Table 7.1 presents example tweets that we annotated as a "good" summary and a "bad" summary on the topic of "Abortion". A good summary tweet is usually self-explanatory; it often contains a phrase that summarizes the event or the situation as well as the author's opinion on it. For example, **"We know it's not okay** `[Indicating a stance] that` **for 40 years politicians have**

denied a woman coverage of abortion [summarizing a situation] just **just because she's poor** [Indicating a stance]".

The author stances are also expressed via certain hashtags that clearly indicate one stance. For example, #BeBoldEndHyde refers to a campaign initiated by an organization "All Above All"[2] to support the termination of the Hyde Amendment, which is a legislative provision that blocked federal funds for abortion services except for a few limited cases and indicates the stance of "pro-choice". #Defundpp is a pro-life stance hashtag supporting several Republican politicans' attempts to defund the organization Planned Parenthood, which has been the largest provider of abortions in the U.S. (Cassata, 2011).

On the other hand, the bad summary tweets are usually not self-explanatory, not well-written, and likely to contain vulgar, informal language. While stances are clear in some of them, the author does not clearly nor logically explain why he/she supports the given stance. Some of them are even off topic.

Based on these observations, we derive three primary components that characterize a good controversy summary tweet as follows:

- **Stance-indicative (S):** A good tweet strongly indicates its stance and is often followed by some particular stance hashtags that are widely used by users from the same stance community. While both good and bad tweets frequently include stance hashtags, the presence of stance hashtags is a positive reinforcement signal if the the quality of tweet is decent.

- **Articulate (A):** A good tweet is clear, persuasive, and logical. It also written with proper language.

[2] https://allaboveall.org/

- **Topically-relevant (T):** A good tweet is relevant and self-explanatory in the context of a particular topic.

7.3.2 Ranking Model

For any controversial topic $\mathcal{T}$, we assume that there are always two stances that are in conflict with each other. We denote these stances as $\mathcal{S}_A$ and $\mathcal{S}_B$. Let Γ be a summary of a given topic $\mathcal{T}$. We let $\Gamma = [\Gamma_A, \Gamma_B]$ that denotes the summary of $\mathcal{S}_A$ and $\mathcal{S}_B$, respectively. We define a model that computes whether a tweet τ is likely to be in the set Γ_A:

$$P(\Gamma_A|\tau) = f(P_S(\mathcal{S}_A|\tau), P_A(\tau), P_T(\tau|\mathcal{T})) \tag{7.1}$$

where $P_S(\mathcal{S}_A|\tau)$ computes how likely a tweet indicates $\mathcal{S}_A$, $P_A(\tau)$ computes how articulate the tweet is, and $P_T(\tau|\mathcal{T})$ computes how relevant the tweet is for the topic.

In the next sections, we discuss how to estimate the first two scores. For the topic relevance score, we use the straightforward probability that the tweet sentence was generated from the language model of the given topic, normalized by the tweet length.

7.4 Estimating Stance-indication

7.4.1 Utility of Hashtags for Stance Detection

In order to generate a stance-aware summary, we first have to identify the stances in each tweet. For stance detection in Tweets, we investigate the utility of "stance hashtags". In Twitter, hashtags are a community-driven convention for adding additional context and metadata to tweets. Given the environment where users are forced to be economical with words due to its 140 character limit, hashtags are often useful, effective, and smart in way that they condense the users' opinion stance or sentiments towards a topic. We observe a certain type of hashtags that are specifically used to express one's opinion on certain issues, which we refer as **stance hashtags**.

Table 7.2: Stance Detection test results.

Method	Abortion	Feminism	Cliamte Change	Atheism	Hillary clinton	Macro F1
ngram (basseline)	0.6106	0.5800	0.4208	0.6394	0.5718	0.5646
hashtag1	0.4580	0.4254	0.2929	0.5455	0.4602	0.4364
hashtag3	0.4409	0.4394	0.3242	0.4875	0.4332	0.4250
hashtag5	0.4522	0.4563	0.3172	0.5165	0.4602	0.4405
hashtag7	0.4007	0.4487	0.3422	0.5468	0.4545	0.4386
hashtag9	0.4304	0.4598	0.3223	0.4944	0.4790	0.4372
hashtag11	0.4406	0.4772	0.3556	0.4813	0.4850	0.4479
hashtag13	0.3911	0.4484	0.3422	0.5115	0.4368	0.4260
hashtag15	0.3965	0.4795	**0.4319**	0.5832	0.4724	0.4727
hashtag17	0.4069	0.4717	0.4208	0.5123	0.4610	0.4545
hashtag19	0.4228	0.4571	0.3256	0.5618	0.4664	0.4467
ngram + hashtag1	**0.6166**	0.5825	0.4208	**0.6419**	0.5718	**0.5667**
ngram + hashtag3	**0.6057**	0.5729	0.4186	**0.6554**	**0.5814**	**0.5668**
ngram + hashtag5	**0.6252**	0.5776	0.4170	**0.6542**	**0.5832**	**0.5714**
ngram + hashtag7	**0.6242**	**0.5879**	0.4180	**0.6542**	**0.5753**	**0.5719**
ngram + hashtag9	**0.6122**	**0.5888**	**0.4219**	**0.6530**	**0.5986**	**0.5749**
ngram + hashtag11	**0.6186**	0.5756	**0.4225**	**0.6665**	**0.6098**	**0.5786**
ngram + hashtag13	0.5950	0.5756	**0.4235**	**0.6489**	**0.6112**	**0.5708**
ngram + hashtag15	0.5960	0.5658	0.4134	**0.6499**	**0.6194**	**0.5689**
ngram + hashtag17	**0.6150**	**0.5846**	0.4186	**0.6494**	**0.6269**	**0.5789**
ngram + hashtag19	**0.6132**	0.5785	0.4173	**0.6458**	**0.6027**	**0.5715**

In SemEval 2016, they released an annotated Twitter dataset with three stances – "favor", "against", and "neutral" – for a given controversial topic for a stance detection task (Mohammad et al., 2016a). In the process of curating this dataset, the organizers explained that they manually curated hashtags to find the candidate tweets in order to annotate a balanced number of tweets from each stance as possible. Several teams that participated in the task reported that they used the manually-curated stance hashtags for their tasks as well.

Hence, we first investigate the utility of hashtags for stance detection. We hypothesize that since certain hashtags serve as user-annotated labels for their stances, relevant hashtags for the tweet will be important signals for stance detection. Hashtags can be viewed as incomplete user annotations in terms of recall. We aim to add the missing relevant hashtags for stance detection.

To find the missing relevant hashtags for the tweets, we train tweet2vec, a character composition model that finds vector space representation of the tweets by learning non-local dependencies in character sequences (Dhingra et al., 2016). Tweet2vec pre-

dicts the hashtags for the given tweets via the learned vector representations. Once we predict the hashtags that the given tweet is likely to be associated with, we use the hashtags as additional or alternative features for stance detection task on Twitter.

In the SemEval 2016 Stance Detection task, while various methods have been submitted, none of the methods outperformed the n-gram baseline that is trained bv SVM classifier. We also train the same SVM classifier to predict the stances of the tweets using only the predicted hashtags and ngrams of the text as well as the predicted hashtags.

Table 7.2 shows the F1 score for each topic and the macro F1 as reported in the competition. Using only hashtags did not outperform the baseline of using ngrams except for one set up in Climate Change, which increased the F1 score by 1% points. When hashtags are used with ngrams, the results were mostly improved. The topic that had the most gain was "Hillary Clinton". In the best case when 17 hashtags were added to the tweet, the F1 score of the stance detection is improved by 5% points. The next topic that had the most gain was "Abortion", which was improve by 1.5% points. In other topics, the gain was about 1% or less. The topics that show a more active stance hashtag usage seemed to benefit more from by the added hashtags as stance context. Both "Abortion" and "Hillary Clinton" are topics that show a high use of stance hashtags because the controversy is related to action-provoking campaigns, such as the one that argues to defund Planned Parenthood (`#defundpp`) or the one that supports voting for Hillary Clinton (`#IamWithHer`) or Donald Trump (`#MAGA`) during the 2016 Presidential Election.

While we have verified that adding relevant hashtags to the tweets provides useful information that helps towards stance detection to some extent, we learned that stance hashtags are particularly helpful keywords for stance detection. Regarding this, the organizers of SemEval 2016 stance detection task stated *"A tweet that has a seemingly favorable hashtag may in fact oppose the target; and this is not uncommon. Similarly*

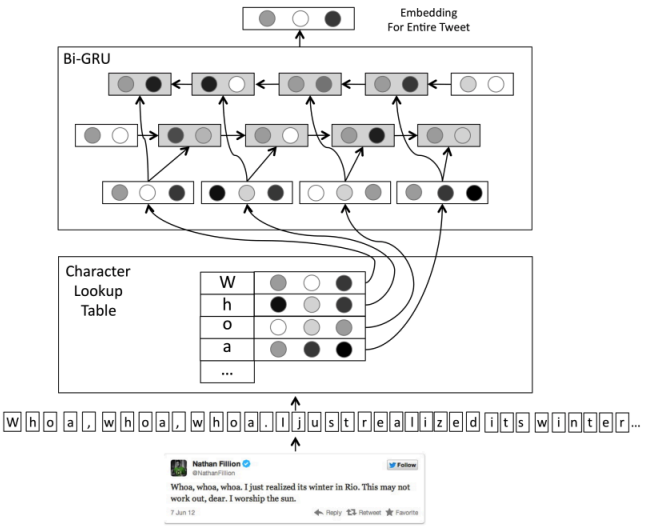

Figure 7.1: Tweet2Vec Model (Dhingra et al., 2016)

unfavorable hashtags may occur in tweets that favor the target", warning that stance hashtags can be easily noisy.

However, we observe that one of the reasons is that stance hashtags that have a seemingly perfectly clear stance are often misused by users on purpose to draw more attention. This is a common practice on Twitter because using popular hashtags are likely to be searched more often hence more visible. Hence, we aim to find stance hashtags that are shown to statistically distinguish the two stance communities, instead of using the manual stance hashtags whose intended stance might be clear but that we do not know how much discrimiative power they actually have for stance detection in the tweets.

7.4.2 Estimating Stance-indication

To estimate stance-indication, we first identify stance hashtags that statistically characterize the stance community. We use the stance hashtags as a proxy to estimate the tweets that indicate the same stance as follows:

$$P_S(\mathcal{S}_A|\tau) = \sum_{h \in \mathcal{H}} P(h|\tau) \cdot P_S(\mathcal{S}_A|h) \cdot P(h)$$

where $\mathcal{H}$ indicates the set of all hashtags and h is a given hashtag. n Then the score boils down to estimating $P(h|\tau)$, a probability that the tweet includes a given hashtag h, and $P_S(\mathcal{S}_A|h)$, a score that indicates how likely it is that h represents $\mathcal{S}_A$. As $\mathcal{S}_A$ and $\mathcal{S}_B$ are mutually exclusive, we penalize ambiguous tweets that are likely to contain stance hashtags of the opposing side by subtracting the score for the opposite stance as follows:

$$P_S(\mathcal{S}_A|\tau) = \sum_{h \in \mathcal{H}_A} \left[P(h|\tau) \cdot P_S(\mathcal{S}_A|h)\right] - \sum_{h \in \mathcal{H}_\mathcal{B}} \left[P(h|\tau) \cdot P_S(\mathcal{S}_B|h)\right]$$

where $\mathcal{H}_A$ and $\mathcal{H}_B$ are the set of stance hashtags that represent $\mathcal{S}_A$ and $\mathcal{S}_B$ respectively.

7.4.3 Identifying Stance Hashtags ($\mathcal{H}_A, \mathcal{H}_B$)

To obtain a set of stance hashtags, we first identify two communities, C_A and C_B, each of which represents two conflicting stances, $\mathcal{S}_A$ and $\mathcal{S}_B$. As introduced by Garimella et al., we construct a user retweet (RT) graph and partition it into two groups (Garimella et al., 2016). We use a simple method that produces only two communities so as not to deal with the extra step of classifying several identified communities to two stances. We leave identifying multiple communities and clustering them into one of the stances of interests to generate the summaries from for the future work.

Once we identify C_A and C_B, we assume that tweets that are written by users from C_A and C_B are likely to indicate $\mathcal{S}_A$ and $\mathcal{S}_B$ respectively. From the two sets of tweets, we compute the information gain (Yang and Pedersen, 1997) that each hashtag gets for the information of the community class when they are present in the tweets: if we know nothing about the tweet but the hashtag presence, which hashtag

best indicates its stance community? Finally, we define $\mathcal{H}_A$, the set of stance hashtag of $\mathcal{S}_A$, as follows.

$$\mathcal{H}_A = \{h \in \mathcal{H} | h \in \mathit{TopN}(IG, \mathcal{H}) \wedge \mathit{freq}_A(h) > \mathit{freq}_B(h)\}$$

where IG is a function that returns the information gain value for the two stance classes for a given hashtag, freq_A is the frequency of h in the tweets published from C_A, and $\mathit{TopN}(IG, \mathcal{H})$ returns the N items that have the highest scores from a given function IG among the items in the given set $\mathcal{H}$. In our experiments, we set $n = 30$, which covers a sufficiently high number of tweets in the community given that the distribution of hashtag frequency follows the power law (Pérez-Melián et al., 2017). We then let $P_S(S_A|h)$ be the normalized score of $IG(h)$ for all hashtags in the set $\mathcal{H}_A$.

7.4.4 Estimating $P(h|\tau)$ via Latent Hashtags

If we think of hashtags as user-generated annotations, hashtags are incomplete annotations. It means that a lack of a certain hashtag does not necessarily mean that it is not a relevant label. To better utilize hashtags as more accurate signals, we make hashtags more complete annotations by estimating $P(h|\tau)$ for all hashtags, the probability that tweet τ generates a hashtag h. Therefore, we adopt a character composition model, Tweet2Vec, which finds a vector space representation of tweets to predict user-annotated hashtags (Dhingra et al., 2016).

By finding the embeddings of tweets and hashtags, we estimate $P(h|\tau)$ for hashtags that were not explicitly used in the given tweet. The model computes the hashtag posterior probability for a given tweet for all hashtags in their softmax layer in order to find the top hashtag predictions. We use this probability as $P(h|\tau)$ for hashtags that were not explicitly used in the given tweet.

Table 7.3: The features used to train a regression model for predicting the level of tweet artriculation.

Feature	Description
Tweet POS Tags (Owoputi et al., 2013)	The ratio of Tweet POS tags
OOV words [3]	The ratio of words that are not in the dictionary
Offensive Words [4]	The ratio of offensive/profane words
POS Tags N-grams	N-grams of Tweet POS Tag sequence
Stop words	The ratio of stop words
Tweet length	The number of characters in a tweet
Avg. word length	The avg. number of characters in tweet words

7.5 Estimating the articulate level

We build a regression model that predicts how well the tweet is written and generate an annotated set of 150 articulate and 150 non-articulate tweets on arbitrary topics. The annotation criteria between the two classes is whether the given tweet is logical, the grammar is sound, and it is written with proper language.

Similarly, Duan et al. propose a classifier to evaluate the content quality of tweets (Duan et al., 2012). In addition to their features, we include a large set of POS tags that are Twitter-specific provided by TweeboParser (Owoputi et al., 2013), N-grams of the POS tags sequence to capture the structural flow of the good sentences, and the ratio of offensive words to penalize usage of inappropriate language, as shown in Table 7.3. This model is generalizable since the features are not content-specific. We trained a logistic regression model and obtained 89.9% classification accuracy using 5-fold cross validation.

7.6 Summary Selection

We propose two algorithms that aggregate the three probability scores to generate the final k summary tweets, which we set as 10 in our experiments. To produce a final summary to equally cover two stances, both algorithms select $k/2$ tweets from each stance.

SumSAT ranks the tweets by setting the aggregation function f (in Eq. 7.1) to be the harmonic mean of the three scores described earlier. HashtagSumSAT, on the other hand, while using the same aggregation function, first identifies the top $k/2$ stance hashtags for each stance and selects the top tweet for each hashtag. While we use the harmonic mean as f, any aggregator can be plugged in. The difference of the two algorithms come from whether it globally ranks the tweets or ranks the tweets per each hashtag.

7.7 Evaluation

We evaluate our methods by running them on real data and conducting user studies to capture the utility of our algorithms.

7.7.1 Experiment Setup

We consider five controversial topics including two short-term, event-based controversies (2016 US Presidential Election and 2017 US National Anthem Protests which we refer to as `#TakeAKnee`), and three long-term ethics-related controversies (Abortion, Feminism, and Climate Change).

Our goal is to generate a summary that can explain why the topic is controversial. For each topic, we generate a pair of summaries and ask 10 participants on Amazon Mechanical Turk which summary better explains the controversy in a double-blind fashion. A pair of summaries were compared twice by two participants. The participants could also say that the quality of the two summaries is the same. To observe whether a subset of tweets whose author's stance is identified from the community generates a better quality summary, we experiment with two cases for each algorithm: (1) using all tweets as summary candidates or (2) using only tweets whose author belongs to one of two stance communities we identified. We distinguish the second case

by adding 'C' (for the community) to the method name. We also generate summaries including the following baseline methods:

- **Random**: A random set of k tweets from a unique set of tweets.
- **MostRT**: The top k most-retweeted tweets in a given day
- **SumBasic (Nenkova and Vanderwende, 2005)**: A general summarization technique. We preprocess the tweets to exclude Twitter-specific stop words. SumBasic algorithm runs as the following:
  - Step 1: for each word w in the input corpus, assign a unigram distribution probability $P(w) = \frac{TF(w)}{|N|}$ where $TF(w)$ is the term frequency of w in the corpus and N is the number of words in the corpus.
  - Step 2: for each sentence S in the corpus, assign the probability by the average of $P(w)$ for all terms w in S.
  - Step 3: pick the highest sentence by the assigned score and add it to the final summary set.
  - Step 4: For each term in the sentence selected from Step 3, reduce the term probability with $P_{new}(w) = P(w) \cdot P(w)$.
  - Step 5: go to Step 2 and repeat until k sentences are chosen.

7.7.2 Results and Discussion

The evaluation shows that our methods were consistently more effective than other baselines across all five topics as shown in Figure 7.2). Overall, SumSAT generated the summaries that were preferred the most (68%) followed by HashtagSumSAT-C (61%). We report the results by the five topics in Figure 7.3.

Controversy summarization as a new task: Overall, both Sumbasic (8%) and Sumbasic-C (42%) generated worse summaries than the naive baselines such as

Table 7.4: The amount of data used to train Tweet2Vec and summary generation. The number in parentheses refers to the number of tweets published by the stance community.

Topic	Tweet2Vec		Summary	
	# Tweets	# Users	# Tweets (# in C)	RT ratio
Election	10.8M	4.3M	10000 (4268)	70.9%
#TakeAKnee	565K	692K	44167 (17217)	71.1%
Abortion	692K	539K	3477 (1262)	57.6%
Feminism	1.7M	1.7M	50323 (20783)	41.3%
Climate Change	546K	360K	10234 (3915)	60.1%

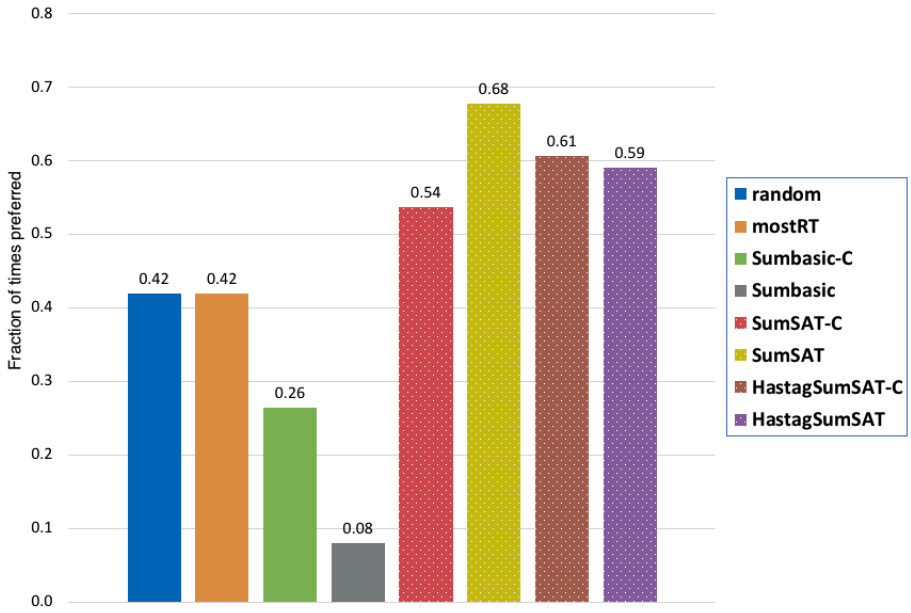

Figure 7.2: The evaluation results by the methods. The rightmost four bars are our methods.

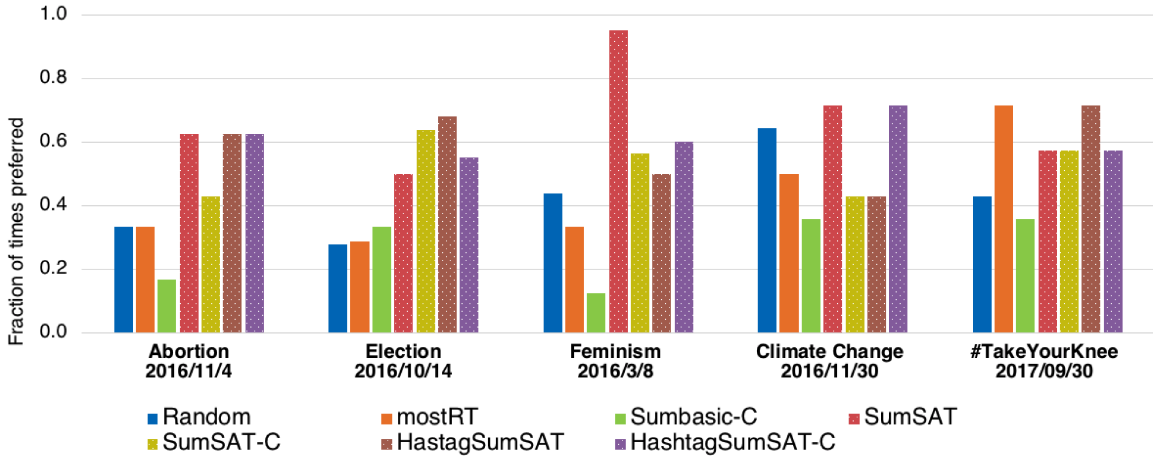

Figure 7.3: The user study results by the topics. The rightmost four bars in each topic are our methods. We did not include SumBasic in the graph because it was the worst method for all topics, being preferred only 8% of times overall.

mostRT or random. This suggests that controversy summarization is an inherently different task from a general topic summarization.

MostRT is often a strong baseline, but its performance is not reliable: For the topic of `#TakeAKnee`, the mostRT baseline was as effective as our top approach. The topic also particularly had a high ratio of retweets compared to other topics (Table 7.4). However, depending on the topic and the day, mostRT can also be the worst feature, even worse than random selection as in the case for the topic of Feminism. For example, the top retweets in Feminism include `'Happy International Women's day!'`. Retweets can often be tweets for entertainment and can easily be dominated by people on one side of stances who are more vocal on Twitter.

Social features seem to be more useful than the content itself in stance summarization: We also learned that in identifying and finding stance-indicative tweets, social features are far more important than the content itself. For example, mostRT outperforms a general summarization technique that only considers the text

content most of the times. This finding aligns with the findings of the previous study on detecting controversy on Twitter (Garimella et al., 2016).

Utility of stance hashtags: While SumSAT was an overall winner, Hashtag-SumSAT outperformed SumSAT for two topics: US Election and #TakeAKnee. We observe a tendency in the event-based controversies like those topics to show more active usage of stance hashtags as there were specific actions people try to promote via stance hashtags. In such type of controversies, stance hashtags were particularly effective to generate a summary around.

7.8 Conclusion

We introduce and tackle a new task of generating a stance-aware summary to explain controversy on social media. Our goal is to provide a tool that helps people navigate controversy effectively. We propose a ranking model that considers three factors that suggest a tweet be part of a good summary derived from our qualitative observations. We assume that a good summary tweet is clear, articulate, and relevant to the topic. Our algorithm characterizes two conflicting stances by identifying two communities from a retweet graph and retrieving the tweets published by them. We define and identify "stance hashtags" that are distinctively used to indicate their opinions in each community and propose a probability model that computes how a tweet is likely to indicate the stance of the community based on the probability that the tweet is likely to generate those hashtags. Our evaluation demonstrates that users prefer the summaries from our methods over the ones from other reasonable baselines.

CHAPTER 8

CONCLUSION AND FUTURE WORK

In this thesis, we studied probabilistic models to identify and explain controversy. In the realm of controversy detection, we argue that the models can be categorized in two types: topic controversy models and document controversy models. Topic controversy models take a topic (i.e., a concept) as a query and output the level of controversy of that topic, whereas document controversy models take a document (i.e., an object) and output the level of controversy for that given document. The two types of model differ in their goal and challenges. Most existing work falls into topic controversy models and implicitly defines controversy as the level of "disputes". Hence, existing work focuses on capturing "disputes" among people within a specific medium, such as Wikipedia and social media. At a high-level, the underlying assumption shared among the existing work is that if people who discuss the given topic display conflicts in some way, the topic is controversial. We argue that many existing topic controversy models fall into a category of a *population-based* topic controversy model, which defines a metric to measure the level of conflict among a group of people that participate in the discussion of the topic. On the other hand, document controversy models have been less studied, particularly from a theoretical modeling perspective. The first part of this thesis investigates the document controversy models.

In Chapter 3, we first developed a probabilistic framework for the controversy detection problem and recast the state-of-the-art algorithm (Dori-Hacohen and Allan, 2015) from that probabilistic perspective. We propose a view that the algorithm is an implementation of an underlying model named kNN-WC. We suggest that kNN-

WC has three properties: (1) P1: kNN-WC has a population-based topic controversy model as a sub-component to estimate the probability of controversy (2) P2: kNN-WC does not directly model non-controversiality (3) P3: the text of a query document does not directly affect the probability of controversiality. The model also suggests that a successful implementation of kNN-WC model would satisfy accurate estimation of two probability components: the probability that a given Wikipedia topic is relevant to the document and the probability that a Wikipedia topic is controversial.

In Chapter 4, we revisited the state-of-the-art algorithm to examine if the algorithm effectively implements the underlying kNN-WC model. We identified two issues with how the probabilities are being estimated in the algorithm. First, while the algorithm generates a single TF10 query from the document to retrieve topics, because documents almost always contain multiple sub-topics, the generated query contains an unknown mixture of different sub-topics and often does not cover all sub-topics properly. Second, while topic controversy models in Wikipedia such as (M score and C score) are used to estimate the probability that a Wikipedia topic is controversial, those scores suffer from sparsity where many specific controversial topics are considered to be non-controversial. Henec, we propose two modifications in the algorithm's framework. The proposed modifications include improving Wikipedia topic retrieval using a text-segmentation based query generation method named TILEQUERY and smoothing controversy scores among topically-related Wikipages for less attended but controversial topics. Our modifications improve the controversy detection classification by 14% more effective in AUC in accuracy.

In Chapter 5, we revisited the three properties, P1, P2 and P3, and hypothesized that those properties might be hindering the model's performance. To test an alternative model that has complementary properties, we propose counter properties P1′, P2′, and P3′, each of which corresponds to the original property. We finally proposed a new document controversy model, Controversy Language Model (CLM).

CLM satisfies the three counter properties by using alternative "language" signals that are obtained from several controversy-indicative signals. By using the language signals, we overcome the sparsity issue that a population-based topic controversy model brought, by transferring the "dispute" signals to "language" that occurred with the disputes (P1′). CLM considers how the probability of controversiality dominates the probability of non-controversiality (P2′). Finally, CLM considers the query document's text directly to estimate the probability that the document is controversial (P3′).

We extensively evaluated the efficacy of CLM by gathering controversial documents from various sources from Wikipedia, news articles, and general Web documents that are retrieved from the controversy-indicative keywords, and the controversy lexicon from previous work. We demonstrated that strongly indicative terms are as helpful for this problem as complicated Wikipedia-based controversy features and more effective than existing lexicons. Our comparative analysis suggests that while kNN-WC is slightly more prone to make false negative errors, CLM is more prone to make false positive errors.

In Chapter 6, we turn to a Wikipedia controversy topic model and point out that existing models do not take *time* into consideration for estimating the probability of controversy. While the existing models are effective at interpreting existing conflict signals into the level of controversy, they are not designed to be adaptive to time. The existing work has used the accumulated edit history as the evidence, some controversy scores such as M score tend to be monotonically increasing over time as more conflicts are included as input. In order to identify controversy that changes over time flexibly, we are in need of a topic controversy model that considers a given time as an input as well as a topic.

As the first straightforward but plausible baseline, we compute a time-window-based M score. Instead of considering accumulated edit history until the query time,

which is the way that has been used in the prior work, we split the edit history and consider only a window of a year to compute M score just for the year. Through a case study, we show that these scores are extremely sparse and most controversial topics follow the same pattern where they only have a few peaks and otherwise appear to be non-controversial. The bigger issue is that once a controversial topic receives a lot of conflicts upon the article creation (if the topic was already controversial before) or the controversy creation, the topic reaches a point to be "matured" or "saturated" that the sign of controversy no longer newly appears. This causes many controversial topics to have low controversy scores in the later years while they are still highly controversial.

Therefore, we distinguish the concept between the observed controversy and the true controversy and argue that the controversy scores that existing topic controversy models estimate are the observed ones and do not always accurately reflect the reality for these reasons. We introduce three models to estimate the true controversy score trend from by interpolating the observed controversy trend and the public interests on the topic. The proposed three models – MCI, ACI, and WCI – compute the true controversy by multiplying the true contention and the true public interests. The three models differ by its way of estimating the true contention. MCI assumes that the true contention is the same as the maximum observed contention until now, ACI as the accumulated level of observed contention, and WCI as the average level of observed contention in the given window of time. We validate our model through a case study and conclude that MCI generates the most reasonable trend especially for long-term controversies while WCI is more adaptive and suitable to predict the controversy trend more accurately for short-term controversial topics.

Finally, in Chapter 7, we pose a new problem of explaining controversy on social media by generating a summary of two conflicting stances by ranking the tweets how likely that a tweet is a representative summary of each stance. We first characterize

three aspects that a good summary tweet should satisfy: a tweet is likely to be part of a good controversy if it (1) indicates a clear stance (2) is articulate and (3) is relevant to the controversial topic of interests. To estimate the probability that a tweet has a clear stance, we first investigate the utility of hashtags in a stance detection task and conclude that enriching the tweet text with k predicted hashtags from tweet embedding improves the accuracy of stance detection task. This suggests that predicted hashtags can be useful features for stance estimation. We use Twitter's retweet network property to first find user stance communities, and extract the stance hashtags that are distinctively used in each community. We finally show that tweets that have semantically close text to the top stance hashtags that best describe the stance community while being articulate and relevant to the topic are more likely to be an effective summary. Our human evaluation shows that our summaries are preferred over other baseline summaries.

8.1 A Theoretical Unifying Perspective on Controversy

While the computational definition of controversy is still an open question in cognitive science, we have attempted to identify the major aspects that contribute to controversy. We previously argued that controversy should be defined and measured with respect to a given population (Jang et al., 2017). In our opinion, we believe that there exists at least five aspects that make up controversy among a given population, namely: contention, popularity, importance, endurance, and conviction. We discuss each aspect, how to capture it, and what existing work has captured.

8.1.1 Contention

Contention generally measures how much dispute the topic has generated among the population, and is probably the most straightforward aspect that make up controversy. Dori-Hacohen (2017) defined it as the ratio of group sizes that hold a conflicting

stance to each other in a way that the level of contention is maximized when the population has split to two equal-sized groups of conflicting stances. Existing work in Wikipedia had slightly different measures to measure the level of disputes among the Wikipedia editors such as the number of terms that have been added and deleted by the editors (Vuong et al., 2008) or the cumulative weighted mutual reverts (Yasseri et al., 2012).

8.1.2 Popularity

Popularity measures how popular the topic is among the given population. When people's interest on the matter is high, things are likely to be easily controversial. Especially in a population-based model, popularity is one of the fundamental aspects that can generate a controversy to begin with. If a topic has no popularity such that no one cares to have an opinion, it would hardly be controversial. We suggest that the popularity can be generally measured by the number of people who show interest in the topic, such as the number of editors who contribute to a Wikipedia article on the given topic, the size of search query volume, or the number of news articles published on the topic.

8.1.3 Importance

Importance signifies how much impact the topic brings to the population in the real world. While importance is a crucial dimension that separates frivolous controversial topics that are highly contentious but do not have any impact in real world such as the well-known "The Dress" or "Yanni vs Laurel controversy" from high-stake controversial topics such as "Brexit" or "2016 US Presidential Election".

While importance itself is difficult to computationally define, in our previous work, we attempted to narrow it down as the number of people that are "affected" by the topic, hence mention the topic in social media (Jang et al., 2017). We denote this sub-population of affected people as Ω_A from a given population Ω. There could

Table 8.1: The number of people who discussed the topic in Wikipedia and Twitter (H2)

	The Dress	Brexit	U.S. Election	Abortion	Toilet paper orientation
# of Wikipedia editors	473	885	2,846	3,152	377
# of Twitter users	286,900	604,100	10,100,000	NA	NA

be various ways to estimate $|\Omega_A|$ depending on how we interpret the meaning of "affected". For example, we suggest three different hypothesis:

- **H1: People who hold a stance on the topic is affected**
- **H2: People who discuss the topic is affected**
- **H3: People who are aware of the topic is affected**

Estimating H3 from News Articles: News reporters are interested in publishing stories that are of interest to the readers. The stories that are worth being published are most likely to be the ones that at least indirectly affect the readers. For example, a local newspaper in Amherst would publish a story that a 30-year-old local Korean restaurant is finally being closed. This story is only of interest to and affects some population in Amherst, and would be less likely to be published by other larger news companies. Therefore, the number of estimated readers of a news article on the topic can be used to approximate $|\Omega_A|$. Let $N_T = \{n_1, n_2, ...n_k\}$ be k relevant news article published on T. Let $View(n)$ be the number of estimated viewers of the news, such as the number of subscribers of the newspaper or the number of users who click on the news.

$$|\Omega_A| = \sum_{i}^{k} View(n_i) \tag{8.1}$$

With lack of access to the information of the $View$ counts, it is practically difficult to compute the value in Eq. 8.1. Instead, we experiment with a simplified assumption where $View(n)$ is always equally k for any n. Although this assumption assumes the

Table 8.2: The number of articles published retrieved by Google News

	The Dress	Brexit	U.S. Election	Abortion	Toilet paper orientation
# of articles returned	1,880	23,500,000	235,000,000	482,000	5,290

same number of k viewers for a local news article and a CNN-featured article, but it relies on the smoothing effect from the number of similar articles published on T if it is originally published by a large newspaper company. Table 1 shows the number of articles returned by Google News on each topic as a preliminary evidence that the number of articles published on more important topics such as "Brexit" and "U.S. Election" are significantly higher than less important topics such as "the Dress" and "Toilet paper orientation" discussion. Here, the topic name itself was used as a query to count the articles published.

However, there are caveats in this definition. The number of views could be affected by the level of popularity. Click-baits headlines constantly strive to increase click views for the news articles. Such factors should be carefully considered not to overuse the measure. Another potential direction to measure importance is to identify the domain of the controversy and have an estimated importance score for each domain. For example, we can assume that any "entertainment" controversy is likely to be less important than any "political" controversy.

8.1.4 Conviction

Conviction looks at how strongly people proclaim their stance. This dimension is motivated that controversy is more heated when people with different stances are more polarized, and each person advocates their stance with stronger voice. This aspect is on how strongly they advocate their own community or attack the other community. We suggest that this can be measured a few different ways as follows:

- **Sentiment in language** A stronger sentiment in the language could signal that users are more convicted with their opinions.

- **The number of vocal users** The number of vocal users who enthusiastically advocate a given stance could be a measure the the conviction in the discussion of the topic. This could be measured by the number of users who use a language with strong sentiment or frequently express their opinions.

- **Network property** Several studies have shown that a controversial topic is likely to generate a divisive community structure on its retweet graph (Conover et al., 2011; Garimella et al., 2016; Fraisier et al., 2017). We could hypothesize that the more exclusively users retweet within their own stance community, the more convicted users are.

- **Polarized usage of language** When the topic is controversial, tweet users are likely to form hashtags that encourage certain movements or agenda, such as `#shoutoutyourabortion` or `#imwithher`. Having such hashtags formed and heavily used in the topic signals that the topic is controversial.

8.1.5 Endurance

Another dimension to consider is "endurance". Cramer previously analyzed the lifespan of controversy cycle: The event first emerges, and it later evolves to a scandal, and to a saga, until it finally stabilizes and is considered to be resolved. Some controversies such as whether abortion should be legalized or climate change is a real concern are long-lived. However, many newly-emerging controversies that are more event-bound have ephemerality, which is an important feature to be captured. Whether the topic has ephemeral pattern in terms of people's attention, and the duration of the controversy signifies the level of the topic controversy.

8.1.6 Summary

We have proposed five aspects for that a topic controversy model would consider. Existing work has captured one or two aspects among them. For example, most

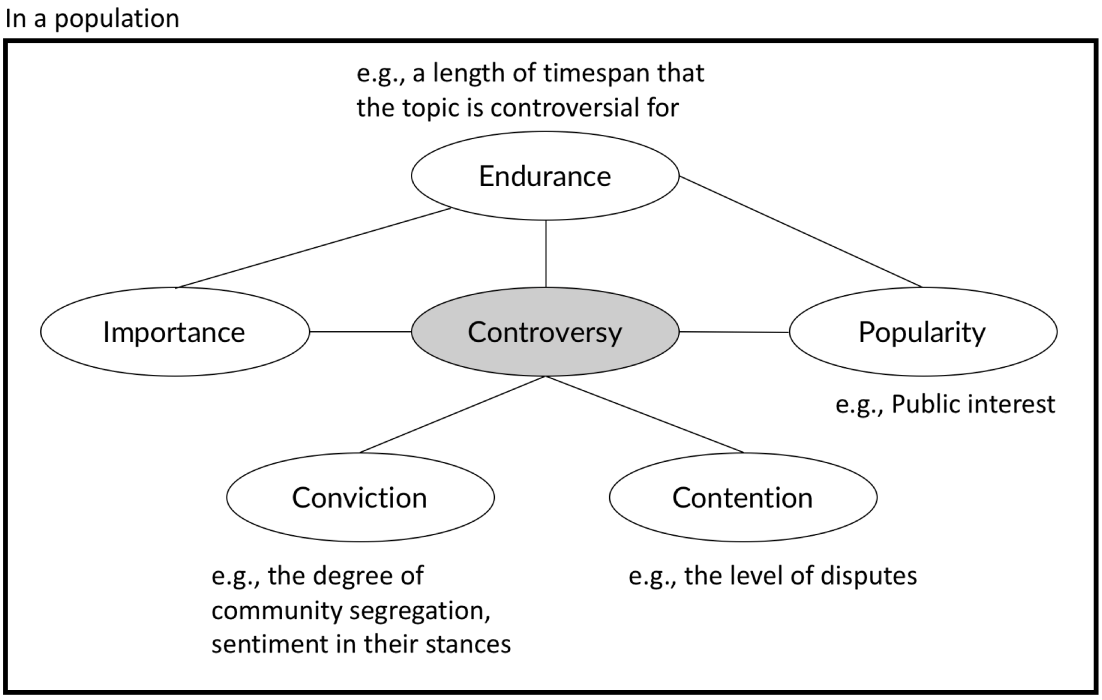

Figure 8.1: A theoretical unifying framework on controversy with five factors that contribute to controversy.

existing Wikipedia controversy detection algorithms have captured the popularity and contention, whereas controversy detection algorithms in social media have captured conviction. In chapter 6, as a model to estimate the true controversy, we have captured popularity and contention. It is questionable whether or not the five aspects should be considered altogether for a complete model as the relationship between these aspects is yet to be investigated. We suspect that the five factors would not completely be independent to each other. For example, "importance" and "popularity" are likely to be correlated to "endurance". We leave this question as future work.

8.2 Future Work

Our work opens up many interesting directions for future work. In Wikipedia, the main signal for controversy is via conflict between two editors, which is captured

via credible editors' activities such as mutual reverts. On the other hand, in social media such as Twitter, people tend to express their opinion rather than by expressing disagreements but by expressing agreements via endorsing other people's opinions (e.g., "retweet" and "like"). The state-of-the-art topic controversy model in social media attempts to capture how the community of one stance is segregated against the other community of the opposite stance. However, despite the characteristics of the different platform that triggers different ways of user involvement, conflicts and segregation could be capturing different aspects of the controversy. For example, while "disputes" can signal how likely the topic is to contain disputable facts and opinions, the degree of "segregation" of the community can signal how strongly people are convinced with their views on the topic with conflicting stances. A unified topic controversy model could be proposed to capture multiple aspects of controversy.

Both the kNN-WC model and CLM utilize Wikipedia topics and their controversy scores. Especially, the kNN-WC model retrieves Wikipedia topics and aggregate the controversy scores of them. However, currently we do not know which sub-topic or portion is particularly controversial of a given topic because the edit history on that page is analyzed as a whole. This makes the controversy detection often too coarse. When a document discusses a certain aspect of a controversial topic that is non-controversial, the document is still highly likely to be classified controversial because our current models do not differentiate that. For example, while 'abortion' is itself a controversial topic, its controversial aspects include political debate and ethical views. Perhaps a document that only discusses the medical procedures or statistical facts may not be controversial, but kNN-WC model would not distinguish the two cases. Therefore, one avenue for addressing this issue is to define and build aspects, or sub-topics of a controversy topic. Identifying specific aspects of the controversy would enable controversy detection at a greater granularity, which will also contribute to generating a useful explanation.

In Chapter 6, we proposed methods to predict the controversy score trend over time. While the methods were validated via a case study, a quantitative evaluation could be designed and conducted to allow us to validate the methods and draw more general conclusions. One task we propose is to perform an extrinsic evaluation in conjunction with CLM by building a time-sensitive CLM drawn from the topics that are controversial in a given year. However, building a dataset that contains the time and controversy judgments would be a tricky problem as annotating the level of controversy retroactively would not be easy.

Lastly, the problem of explaining controversy is still at its early stage and we hope that our work in Chapter 7 brings more attention to this problem in the future. This problem can be extended in many ways. The current method is limited in that it utilizes hashtags to estimate the stance of a tweet. Because not all controversial topics have developed stance hashtags, the method is less effective if the given topic does not have prominent stance hashtags. As the controversial topic dynamically changes and gets updated, an effective method for a temporal summary from social media can be investigated.

APPENDIX

A LIST OF TOP 250 WIKIPEDIA ARTICLES THAT ARE USED FOR CLM

Table A.1: A sample long table.

Rank	Wikipedia Title
1	Antinomian Controversy
2	Teach the Controversy
3	Controversy (law)
4	Scientific controversy
5	Recent history of the District of Columbia Fire and Emergency Medical Services Department
6	Lordship salvation controversy
7	Chicago & Northwestern R. Co. v. Crane
8	Shubhodaya Controversy
9	Vaccine controversies
10	Socinian controversy
11	Nature fakers controversy
12	List of American television episodes with LGBT themes, 1990– 1997
13	Free Grace theology
14	Hillary: The Movie
15	Lars Vilks Muhammad drawings controversy
16	Controversy
17	Investiture Controversy
18	Darwinism, Design and Public Education
19	Rape and pregnancy controversies in United States elections, 2012
20	American Presbyterianism
21	Concerns and controversies at the 2008 Summer Olympics
22	Discovery Institute
23	Intelligent design movement
24	Goguryeo controversies
25	Christmas controversy
26	Amazon.com controversies

Continued on next page

Table A.1 – continued from previous page

Rank	Wikipedia Title
27	Controversy over the use of Manchester Cathedral in Resistance: Fall of Man
28	Telecoms Package
29	John Wilson (minister)
30	Ian Meckiff
31	Luis de Molina
32	Opinions on the Jyllands–Muhammad cartoons controversy
33	Al Qa'qaa high explosives timeline
34	Joseph Desha
35	List of Australian sports controversies
36	Arian controversy
37	American Idol controversies
38	Controversy and Other Essays in Journalism
39	Vestments controversy
40	Transfermium Wars
41	Osiandrian controversy
42	The Cartoons that Shook the World
43	Intelligent design and science
44	David Levine (medical administrator)
45	List of chemical elements naming controversies
46	Scouting controversy and conflict
47	Dungeons & Dragons controversies
48	Simon Fraser University 1997 harassment controversy
49	Singur Tata Nano controversy
50	International reactions to the Jyllands–Muhammad cartoons controversy
51	Sexuality (Prince song)
52	Archpriest Controversy
53	Boom Shaka
54	Riverside Park Management
55	Vea
56	Ako Controversy
57	UBS tax evasion controversy
58	California textbook controversy over Hindu history
59	Possibilism (geography)
60	Chief Illiniwek
61	Illinois High School Association
62	Japanese history textbook controversies
63	Cooks Source infringement controversy
64	The Wikipedia Revolution
65	Limited appearance

Continued on next page

Table A.1 – continued from previous page

Rank	Wikipedia Title
66	Betty Granger
67	Wildlife Protection Act of 2010
68	Inul Daratista
69	Cambridge capital controversy
70	Bye Bye (TV series)
71	Bangorian Controversy
72	Academic freedom at Brigham Young University
73	Institute for Canadian Values ad controversy
74	Fundamentalist – Modernist Controversy
75	Paul Aussaresses
76	Old Court – New Court controversy
77	Kathryn Lindskoog
78	Hindmarsh Island bridge controversy
79	Timeline of plesiosaur research
80	John O'Donoghue expenses controversy
81	Hockey stick controversy
82	The Panda's Thumb (blog)
83	Bosom Friends affair
84	Julius Micrander
85	Influence of Sesame Street
86	Hawaii State District Courts
87	Campe (poem)
88	The Great Controversy (book)
89	Abbey Mills Mosque
90	Half Pint Brawlers
91	Murray Deaker
92	DADVSI
93	History of the hamburger
94	The Nightingale casting controversy
95	Cannibal film
96	Vierordt's law
97	Dismissal of U.S. attorneys controversy
98	Ellen G. White
99	Macaca (term)
100	Climatic Research Unit email controversy
101	Evangelical Lutherans in Mission
102	Capitol Loop
103	Baya al Ward
104	Brown Dog affair
105	James of Brescia
106	Brian Alters
107	Steven Courtney

Continued on next page

Table A.1 – continued from previous page

Rank	Wikipedia Title
108	Ferenc GyurcsÃąny plagiarism controversy
109	Inger Louise Valle
110	Antarctica cooling controversy
111	Thomas Cornell (settler)
112	Meletius of Lycopolis
113	Gerald Graff
114	Anglo Irish Bank hidden loans controversy
115	Second Test, 2007âĂŞ08 BorderâĂŞGavaskar Trophy
116	Donald Gordon (Canadian businessman)
117	Sheldon v. Sill
118	Zsolt SemjÃľn academic misconduct controversy
119	W. A. C. Bennett Dam
120	Marcela AcuÃśa
121	Edward Einhorn
122	Molecular assembler
123	Sweden in the Eurovision Song Contest 2006
124	Employee stock option
125	Controversies surrounding Yasukuni Shrine
126	Joachim Westphal (of Hamburg)
127	Valentin Ernst LÃűscher
128	John Cotton (minister)
129	Wayne Laugesen
130	Jyllands– Posten Muhammad cartoons controversy
131	41st Academy Awards
132	John C. Browne
133	Erhardt v. Boaro, (113 U.S. 527)
134	The Holy Virgin Mary
135	Derek Freeman
136	War of the Theatres
137	Fuda Cancer Hospital–Guangzhou
138	Kikuyu controversy
139	Rotvoll controversy
140	Controversy of Nanzhao
141	Controversy Tour
142	Alta controversy
143	Pichilemu political controversies
144	Texas Instruments signing key controversy
145	Apple and Adobe Flash controversy
146	National Football League controversies
147	DelisleâĂŞRichler controversy
148	Controversies of the United States Senate election in Virginia, 2006
149	DelisleâĂŞRichler controversy

Continued on next page

Table A.1 – continued from previous page

Rank	Wikipedia Title
150	Frank C. Hibben
151	List of controversial album art
152	Manufactured controversy
153	Thomas William Marshall
154	Summer reading program
155	Sarawak Tribune
156	Becket controversy
157	Controversy (song)
158	Easter controversy
159	2012 Karnataka video clip controversy
160	Calvin Butler Hulbert
161	Exxon Mobil Corp. v. Allapattah Services, Inc.
162	Alan Bean (activist)
163	Immunization Alliance
164	Sectarian violence In Pakistan (1988)
165	Amir Taheri
166	DePauw University Delta Zeta discrimination controversy
167	List of Internal Revenue Service political profiling controversies
168	Tax controversy
169	Chester's guide to: The controversy
170	Samuel Fancourt
171	Heather Bresch M.B.A. controversy
172	Vestment
173	Pinot noir passing– off controversy
174	Wikipediocracy
175	Three– Chapter Controversy
176	Jan Esper
177	History of the EastâĂŞWest Schism
178	History of Eastern Orthodox Christian theology
179	Stem cell controversy
180	Trijicon biblical verses controversy
181	Hassi Messaoud mob attacks against women
182	Old SideâĂŞNew Side Controversy
183	George W. Bush military service controversy
184	Rod Blagojevich controversies
185	Tantri controversy
186	Olympic Games scandals and controversies
187	Contents of the United States diplomatic cables leak (Indonesia)
188	Florida Circuit Courts
189	High School Stories
190	James D. Bales
191	Renaissance Unity Interfaith Spiritual Fellowship

Continued on next page

Table A.1 – continued from previous page

Rank	Wikipedia Title
192	MVDDS dispute
193	Ahmed Akkari
194	Pat Buchanan presidential campaign, 2000
195	Definitions of abortion
196	Elisha Gray and Alexander Bell telephone controversy
197	Matt Sanchez
198	Kunicon
199	Gola River
200	Paradise Hotel (Hyderabad)
201	Controversies surrounding Silvio Berlusconi
202	Coma White
203	Scientology and psychiatry
204	HGH controversies
205	He Liked to Feel It
206	Mapping controversies
207	Beginning of pregnancy controversy
208	Asmachta (Talmudical hermeneutics)
209	2004 NCAA Division I–football season
210	Truth in Science
211	Let's Work
212	2013 Senate of the Philippines funds controversy
213	Sault Ste. Marie language resolution
214	Richard Deth
215	Local Church controversies
216	Controversy and criticism of The Voice of the Philippines
217	1960 English cricket season
218	Political views of Paul Robeson
219	J. Krishna Palemar
220	Kathavatthu
221	Manitoba Public Schools Act
222	Game Rating Board
223	Brigitte BarÃĺges
224	Mohamed El Naschie
225	Brigitte BarÃĺges
226	Federal Vision
227	1921 NFL Championship controversy
228	Nkandla (homestead)
229	Controversies in autism
230	Tata Tapes controversy
231	Lipid hypothesis
232	Gilles Bourdouleix
233	Jytte Klausen

Continued on next page

Table A.1 – continued from previous page

Rank	Wikipedia Title
234	Pasquill (the Cavaliero)
235	Stephen Patrington
236	Hull Council election, 1998
237	Godless (novel)
238	Per Edgar Kokkvold
239	Joe Horn shooting controversy
240	Language of adoption
241	Karmapa

BIBLIOGRAPHY

Talos targets disinformation with fake news challenge victory., 2017. http://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html, accessed: 2019-02-01.

J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 10–18, New York, NY, USA, 2001. ACM.

H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

J. Anderson and L. Raine. The future of truth and misinformation online. *Pew Research Center*, 2017. http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online, accessed: 2018-10-05.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885. Association for Computational Linguistics, 2016.

R. Awadallah, M. Ramanath, and G. 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.

K. Beelen, E. Kanoulas, and B. van de Velde. Detecting controversies in online news media. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1069–1072, 2017.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2001.

U. Brandes, P. Kenis, J. Lerner, and D. 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.

E. K. Brunson. The impact of social networks on parents' vaccination decisions. *Pediatrics*, 131:5, 2013.

J. Callan and M. Hoy. Clueweb09 data set, 2009. URL `http://boston.lti.cs.cmu.edu/Data/clueweb09/`. accessed: 2019-01-08.

M.-A. Cartright, E. Aktolga, and J. Dalton. Characterizing the subjectivity of topics. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 642–643, New York, NY, USA, 2009. ACM.

D. Cassata. Planned parenthood, abortion and the budget fight. *The Seattle Times*, April 2011. `https://www.seattletimes.com/nation-world/planned-parenthood-abortion-and-the-budget-fight`, accessed: 2018-01-19.

D. Chakrabarti and K. Punera. Event summarization using tweets. In *The Sixth International AAAI Conference on Weblogs and Social Media*, ICSWM '11, 2011.

Y. Choi, Y. Jung, and S.-H. Myaeng. Identifying controversial issues and their sub-topics in news articles. In *Intelligence and Security Informatics*. Springer, 2010.

M. Coletto, V. R. K. Garimella, A. Gionis, and C. Lucchese. A motif-based approach for identifying controversy. In *ICWSM '17*, Proceedings of the 11th International Conference on Web and Social Media, pages 496–499, 2017.

M. Conover, J. Ratkiewicz, M. R. Francisco, B. Goncalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *The 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 89–96, 2011.

D. Corney, D. Albakour, M. Martinez, and S. Moussa. What do a million news articles look like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016)*, NewsIR '16, 2016.

P. A. Cramer. *Controversy as news discourse*, volume 19. Springer Science & Business Media, 2011.

S. Das, A. Lavoie, and M. Magdon-Ismail. Manipulation among the arbiters of collective intelligence: How wikipedia administrators mold public opinion. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, CIKM '13, pages 1097–1106, 2013.

S. Denyer. Confusion over north koreaâĂŹs definition of denuclearization clouds talks, 2018. URL `https://www.washingtonpost.com/world/asia_pacific/confusion-over-north-koreas-definition-of-denuclearization-clouds-talks/2019/01/15/c6ac31a8-16fc-11e9-a896-f104373c7ffd_story.html?noredirect=on&utm_term=.1ca81f442769`. accessed: 2019-01-16.

A. Depalma. Bre-x: From rags to riches, back to rags. *The New York Times*, 1997. https://www.nytimes.com/1997/05/06/business/bre-x-from-rags-to-riches-back-to-rags.html, accessed: 2018-11-15.

C. Dewey. Facebook fake-news writer: I think donald trump is in the white house because of me. *The Washington Post*, 2016. https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-isin-the-white-house-because-of-me, accessed: 2018-10-06.

B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, and W. W. Cohen. Tweet2vec: Character-based distributed representations for social media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 269–274, 2016.

S. Dori-Hacohen. *Controversy Analysis and Detection*. PhD thesis, University of Massachusetts, September 2017.

S. Dori-Hacohen and J. Allan. Detecting controversy on the web. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1845–1848, New York, NY, USA, 2013. ACM.

S. Dori-Hacohen and J. Allan. Automated controversy detection on the web. In *In Proceedings of the 37th European Conference on IR Research on Advances in Information Retrieval*, ECIR '15, pages 423–434, 2015.

S. Dori-Hacohen, E. Yom-Tov, and J. Allan. Navigating controversy as a complex search task. In *Proceedings of the First International Workshop on Supporting Complex Search Tasks co-located with the 37th European Conference on Information Retrieval*, SCST@ECIR, 2015.

S. Dori-Hacohen, D. Jensen, and J. Allan. Controversy detection in wikipedia using collective classification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 797–800, New York, NY, USA, 2016. ACM.

J. Douglas. Fake news: improved critical literacy skills are key to telling fact from fiction. *The Guardian*, 2017. https://www.theguardian.com/teacher-network/2017/oct/17/fake-news-improved-critical-literacy-skills-teaching-young-people, accessed: 2018-10-06.

Y. Duan, Z. Chen, F. Wei, M. Zhou, and H. Shum. Twitter topic summarization by ranking tweets using social influence and content quality. In *26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, COLING '12, pages 763–780, 2012.

K. El-Arini and C. Guestrin. Beyond keyword search: Discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 439–447, 2011.

O. Fraisier, G. Cabanac, Y. Pitarch, R. Besançon, and M. Boughanem. Uncovering like-minded political communities on twitter. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 261–264, 2017.

D. Ganguly, J. Leveling, and G. J. F. Jones. Automatic generation of query sessions using text segmentation. In *ECIR 2011 Workshop on Information Retrieval Over Query Sessions*, 2011a.

D. Ganguly, J. Leveling, W. Magdy, and G. J. F. Jones. Patent query reduction using pseudo relevance feedback. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1953–1956. ACM, 2011b.

K. Garimella. *Polarization on Social Media*. PhD thesis, Aalto University, February 2018.

K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 33–42. ACM, 2016.

K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 81–90, New York, NY, USA, 2017.

S. L. Gerhart. Do web search engines suppress controversy? *First Monday*, 9(1), 2004. `http://firstmonday.org/issues/issue5_2/choo/index.html`, accessed: 2018-10-06.

J. Gottfried and E. Shearer. News use across social media platforms 2016. *Pew Research Center*, May 2016. `http://www.journalism.org/2016/05/26/news-use-acrosssocial-media-platforms-2016`, accessed: 2018-10-06.

P. H. C. Guerra, W. Meira, C. Cardie, and R. D. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *The Seventh International AAAI Conference on Weblogs and Social Media*, ICSWM '13, 2013.

A. Guess, B. Nyhan, and J. Reifler. Selective exposure to disinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, 2018.

J. Guo, Y. Lu, T. Mori, and C. Blake. Expert-guided contrastive opinion summarization for controversial issues. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1105–1110, New York, NY, USA, 2015. ACM.

K. Gyllstrom and M.-F. Moens. Clash of the typings - finding controversies and children's topics within queries. In *ECIR*, 2011.

A. Hanselowski, A. P. V. S., B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych. A retrospective analysis of the fake news challenge stance detection task. *CoRR*, abs/1806.05180, 2018. URL `http://arxiv.org/abs/1806.05180`.

H. Haselgrove. Wikipedia page-to-page link database, 2009. URL `http://haselgrove.id.au/wikipedia`.

M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, Mar. 1997.

M. Ingram. Here's what's wrong with algorithmic filtering on twitter, 2016. `http://fortune.com/2016/02/08/twitter-algorithm`.

D. I. Inouye and J. K. Kalita. Comparing twitter summarization algorithms for multiple post summaries. *PASSAT and SocialCom*, pages 298–306, 2011.

J. Jackson. Twitter accounts really are echo chambers, study finds. *The Guardian*, 2017. `https://www.theguardian.com/politics/2017/feb/04/twitter-accounts-really-are-echo-chambers-study-finds`, accessed: 2018-10-06.

M. Jang and J. Allan. Improving automated controversy detection on the web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 865–868, New York, NY, USA, 2016. ACM.

M. Jang and J. Allan. Explaining controversy on social media via stance summarization. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, pages 1221–1224, New York, NY, USA, 2018. ACM.

M. Jang, J. Foley, S. Dori-Hacohen, and J. Allan. Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2069–2072, New York, NY, USA, 2016. ACM.

M. Jang, S. Dori-Hacohen, and J. Allan. Modeling controversy within populations. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 141–149, New York, NY, USA, 2017. ACM.

K. A. Johnson and D. Goldwasser. Identifying stance by analyzing political discourse on twitter. In *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*, pages 66–75, 2016.

G. Karypis and V. Kumar. *METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices*, September 1998.

Y. Kim. *Searching Based on Query Documents*. PhD thesis, University of Massachusetts, 2014.

Y. Kim and W. B. Croft. Diversifying query suggestions based on query documents. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 891–894, New York, NY, USA, 2014. ACM.

A. Kittur, B. Suh, B. A. Pendleton, and E. 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.

E. Kolbert. Why facts don't change our minds. *The New Yorker*, 2017. https://www.newyorker.com/magazine/2017/02/27/why-facts-dont-change-our-minds, accessed: 2018-10-06.

I. Lapowsky. In a Fake Fact Era, Schools Teach the ABCs of News Literacy. *Wired*, 2017. https://www.wired.com/2017/06/fake-fact-era-schools-teach-abcs-news-literacy/, accessed: 2018-10-03.

C.-J. Lee and W. B. Croft. Generating queries from user-selected text. In *Proceedings of the 4th Information Interaction in Context Symposium on - IIIX '12*, pages 100–109, 2012.

J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195, 2015.

D. Leonhardt. Can't grasp credit crisis? join the club. *The New York Times*, 2008. https://www.nytimes.com/2008/03/19/business/19leonhardt.html, accessed: 2018-10-04.

Q. McNemer. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:2:153–157, 1947.

Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo. Controversy and sentiment in online news. In *Computation+Journalism Symposium 2014*, 2014.

A. Mitchell, J. Gottfried, M. Barthel, and E. Shearer. Pathways to news. *Pew Research Center*, 2016. http://www.journalism.org/2016/07/07/pathways-to-news, accessed: 2018-10-06.

S. Mohammad, S. Kiritchenko, P. Sobhani, X.-D. Zhu, and C. Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *SemEval@NAACL-HLT*, 2016a.

S. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *CoRR*, 2016b.

S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. Semeval-2016 task 6: Detecting stance in tweets. SemEval, June 2016c.

A. Nenkova and L. Vanderwende. The impact of frequency on summarization. Technical report, Microsoft Research, 2005.

A. Nusca. *Fortune*, 2017. http://fortune.com/2017/12/28/apple-iphone-battery-apology/, accessed: 2019-01-15.

O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *HLT-NAACL*, pages 380–390, 2013.

P. Pannaraj. Us measles outbreaks catalyzed by vaccine hesitancy. *Infectionous Disease in Children*, 2018. https://www.healio.com/pediatrics/vaccine-preventable-diseases/news/print/infectious-diseases-in-children/%7B8073077c-43e9-407a-8766-4752884bb162%7D/us-measles-outbreaks-catalyzed-by-vaccine-hesitancy, accessed: 2018-09-27.

E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press HC, 2011.

H. J. Parkinson. Click and elect: How fake news helped donald trump win a real election. *The Guardian*, 2016. https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies, accessed: 2018-10-02.

M. J. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 66–76, Stroudsburg, PA, USA, 2010.

J. A. Pérez-Melián, J. A. Conejero, and C. Ferri. Zipf's and benford's laws in twitter hashtags. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pages 84–93, 2017.

D. Pomerleau and D. Rao. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake News Challenge*, 2017. http://www.fakenewschallenge.org/, accessed: 2019-02-01.

A.-M. Popescu and M. 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.

H. S. Rad and D. 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.

U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76(3):036106, 2007.

M. Read. Donald trump won because of facebook. *New York Magazine*, 2016. http://nymag.com/selectall/2016/11/donald-trump-won-because-of-facebook.html, accessed: 2018-10-06.

M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659. The Association for Computer Linguistics, 2013.

Reddit. Study shows most experts think iq differences between races are due to genetic-evolutionary factors, what does it mean?, 2018a. URL https://www.reddit.com/r/samharris/comments/88yj3p/https://www.reddit.com/r/AskAnthropology/comments/6y1fta/study_shows_most_experts_think_iq_differences/. accessed: 2019-01-15.

Reddit. Survey of expert opinion on intelligence: Causes of international differences in cognitive ability tests, 2018b. URL https://www.reddit.com/r/samharris/comments/88yj3p/survey_of_expert_opinion_on_intelligence_causes/. accessed: 2019-01-15.

Reddit. What experts in the field think about national differences in iq and academic performance, 2018c. URL https://www.reddit.com/r/samharris/comments/683z84/what_experts_in_the_field_think_about_national/. accessed: 2019-01-15.

H. Rindermann, D. Becker, and T. R. Coyle. Survey of expert opinion on intelligence: Causes of international differences in cognitive ability tests. 2016.

X. N. T. Robin, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Muller. Package 'proc', 2014. (http://cran.r-project.org/web/packages/pROC/pROC.pdf).

H. Roitman, S. Hummel, E. Rabinovich, B. Sznajder, N. Slonim, and E. Aharoni. On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 991–996, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

S. M. Rose, Tony and M. Whitehead. The Reuters Corpus Volume 1. In *LREC 2002*, 2002.

M. Rosvall and D. Axelsson. The map equation. *The European Physical Journal Special Topics*, 178:13–23, 2009.

A. K. Ryan Mac, Charlie Warzel. Growth at any cost: Top facebook executive defended data collection in 2016 memo âĂŤ and warned that facebook could get people killed. *Buzzfeed News*, 2019. https://www.buzzfeednews.com/article/ryanmac/growth-at-any-cost-top-facebook-executive-defended-data, accessed: 2019-02-01.

E. Sandhaus. The New York Times annotated corpus. *LDC*, 6(12):e26752, 2008.

H. Sepehri Rad, A. Makazhanov, D. Rafiei, and D. 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. ISBN 978-1-4503-1335-3.

D. Shahaf and C. Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 623–632, New York, NY, USA, 2010. ACM.

B. Sharifi, M.-A. Hutton, and J. Kalita. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 685–688, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

C. Silverman. This analysis shows how fake election news stories outperformed real news on facebook. *Buzzfeed News*, 2016. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook, accessed: 2018-10-02.

C. Silverman and J. Singer-Vine. Most americans who see fake news believe it, new survey says. *BuzzFeed News*, 2016. https://www.buzzfeednews.com/article/craigsilverman/fake-news-survey, accessed: 2018-10-05.

M. D. Smucker and J. Allan. Find-similar: Similarity browsing as a search tool. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 461–468, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148250. URL http://doi.acm.org/10.1145/1148170.1148250.

M. Snyderman and S. Rothman. *The IQ Controversy, the Media and Public Policy*. New Brunswick, NJ: Transaction., 1988.

W. Su, Y. Yuan, and M. Zhu. A relationship between the average precision and the area under the roc curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 349–352, New York, NY, USA, 2015. ACM.

R. R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertész. Edit wars in Wikipedia. *PASSAT & SocialCom*, pages 724–727, 2011.

L. Vogel. Viral misinformation threatens public health. *Canadian Medical Association Journal*, 189(50):Article E1567, 2017.

B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, H. W. Lauw, and K. 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.

Wikipedia. Critical literacy, 2019a. URL `https://en.wikipedia.org/wiki/Critical_literacy`. accessed: 2019-01-03.

Wikipedia. Kim jong-il, 2019b. URL `https://en.wikipedia.org/wiki/Kim_Jong-il`. accessed: 2019-01-10.

E. Wiseman. Get out of my echo chamber. it's cosy in here. *The Guardian*, 2016. `https://www.theguardian.com/lifeandstyle/2016/dec/11/get-out-of-my-echo-chamber-its-cosy-in-here`, accessed: 2018-09-25.

Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, ICML '97, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

Y. Yang, N. Bansal, W. Dakka, P. Ipeirotis, N. Koudas, and D. Papadias. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 34–43, New York, NY, USA, 2009. ACM.

T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész. Dynamics of conflicts in wikipedia. *PloS one*, 7(6):e38869, 2012.

E. Yulianti, S. Huspi, and M. Sanderson. Tweet-biased summarization. *J. Assoc. Inf. Sci. Technol.*, 67(6):1289–1300, June 2016.

K. Zielinski, R. Nielek, A. Wierzbicki, and A. Jatowt. Computing controversy: Formal model and algorithms for detecting controversy on wikipedia and in search queries. *Inf. Process. Manage.*, 54:14–36, 2018.