Abstract

The thesis examines political discourse quality online and proposes a methodology for analyzing online conversations in an automated way. The study builds on Habermas’ work by examining the quality of the public sphere in a digital age. Primarily, it examines the portion of the public sphere which deals with political discussions on online platforms. The proposed technique, DelibAnalysis, is a combination of random forests classification and k-means clustering using term-frequency inverse-document-frequency. The DelibAnalysis methodology is applied to a diverse dataset of online conversations between citizens and elected representatives in Canada, the United States and the United Kingdom using Facebook and blog platforms. This analysis is used to derive insights about the state of the online public sphere and the differences between platforms and discussion frameworks. The objective of this research is to provide a systematic framework for the semi-automated discourse quality analysis of large datasets, and in applying this framework, to yield insight into the structure and features of political discussions online.

Reference


DOI : 10.13097/archive-ouverte/unige:112458
URN : urn:nbn:ch:unige-1124585
DelibAnalysis: Understanding online deliberation through automated discourse quality analysis and topic modeling

Doctoral Dissertation
June 2018

Eleonore Fournier-Tombs
Université de Genève
Faculté des sciences de la société
Doctorante en sciences de la société mention système informatiques
eleonore.fournier-tombs@etu.unige.ch
Peux ce que veux. Allons-y.

Roméo Dallaire
Acknowledgements

This thesis would not have been possible without the steady and insightful support of Professor Giovanna Di Marzo Serugendo, who not only believed in this subject from the very beginning, but also allowed me great independence and flexibility to work around a complicated work and family schedule. I also want to thank my partner, Curtis, who saw me through my Masters thesis, encouraged me to take on the PhD and patiently listened to the ups and downs of excitement and doubts. Of course, I cannot avoid thanking my little one, Julien, who was born during this process and who has slept through the night from the beginning and inspired me to become a better version of myself.

I also want to thank my incredible colleagues, both at the United Nations and at RedOwl Analytics, for motivating me to finish my dissertation and for teaching me some tricks of the trade in computational analysis. I learned a lot of skills at work that I was able to apply here.

I would also like to thank the members of my defense committee – Jean-Henry Morin, André Bächiger, Paola Merlo, Rhon Teruelle and Gilles Falquet, for their insightful comments and questions, opening up for me the future after my dissertation.

I want to thank my family and friends, for always being supportive and excited about this project, in particular my parents George and Angéline and my siblings, Nathanael, Iona and Myriam.

Finally, I dedicate this to my two grandmothers, Eleanor Jean and Marie-Dominique, each of whom, in their own way, taught me the power of critical thinking, ambition and fearlessness.
Abstract

The following thesis examines political discourse quality online and proposes a methodology for analyzing online conversations in an automated way. The study is based on seminal work on the Discourse Quality Index (DQI) by Steenbergen et al (2003) and is primarily grounded in Habermas’ *The Structural Transformation of the Public Sphere* (1961). The study builds on Habermas’ work by examining the quality of the public sphere in a digital age. Primarily, we examine the portion of the public sphere which deals with political discussions on online platforms. Drawing on the DQI, we manually code a portion of political comments to create a training dataset for a classifier, which can then be used on much larger data corpuses. We propose a topic clustering algorithm which gives context to the DQI score for any given conversation.

The proposed technique, which we call DelibAnalysis, is a combination of random forests classification and k-means clustering using term-frequency inverse-document-frequency. This methodology was selected after comparative testing using logistic regression and support vector machines (SVM), as well as a variety of data formats and parameters.

After presenting the DelibAnalysis methodology, we apply it to a diverse dataset of online conversations between citizens and elected representatives in Canada, the United States and the United Kingdom using Facebook and blog platforms. We use this analysis to derive insights about the state of the online public sphere and the differences between platforms and discussion frameworks.

The objective of this research is to provide a systematic framework for the semi-automated discourse quality analysis of large datasets, and in applying this framework, to yield insight into the structure and features of political discussions online.
Table of Contents

Acknowledgements ........................................................................................................................................ iii
Abstract ...................................................................................................................................................... iv
Table of Contents .......................................................................................................................................... i
Figures ........................................................................................................................................................... v
Tables ............................................................................................................................................................. vi
1. Introduction .............................................................................................................................................. 1
   1.1 Methodology and approach .................................................................................................................. 3
      1.1.1 Reliability and validity of research .............................................................................................. 6
   1.2 Contributions ....................................................................................................................................... 8
   1.3 Research ethics ...................................................................................................................................... 8
   1.4 Current philosophical debates ........................................................................................................... 9
   1.5 Structure of the document .................................................................................................................... 11
   1.6 Summary ............................................................................................................................................. 12
2. State of the Art ......................................................................................................................................... 13
   2.1 Political Discourse and Deliberation ................................................................................................. 13
   2.2 The Discourse Quality Index ............................................................................................................ 15
   2.3 Network Analysis ............................................................................................................................... 18
   2.4 Automated Discourse Quality Analysis ........................................................................................... 22
   2.5 Conclusions from state of the art section ......................................................................................... 25
   2.6 Summary ........................................................................................................................................... 25
3. DelibAnalysis Framework ......................................................................................................................... 26
   3.1 Collection and Processing .................................................................................................................... 27
      3.1.1 Download data ................................................................................................................................. 27
      3.1.2 Process data .................................................................................................................................... 28
      3.1.3 Adding features such as character count and platform ............................................................... 28
   3.2 Manual Labelling ................................................................................................................................. 29
      3.2.1 Selecting a portion of the dataset for manual labelling ................................................................ 30
      3.2.2 Adding the score for each of the six DQI indicators ................................................................... 30
      3.2.3 Combining indicator scores to obtain the DQI score ................................................................ 30
Chapter 1. Introduction

3.2.4 Converting the DQI score into a DQI category .................................................. 30
3.3 Classifier Training ..................................................................................................... 31
  3.3.1 Split the labelled dataset into training and testing sets ....................................... 31
  3.3.2 Vectorize words in data ....................................................................................... 31
  3.3.3 Create random forests object ............................................................................ 31
  3.3.4 Tweak classifier parameters .............................................................................. 32
  3.3.5 Fit training set to random forests object ............................................................ 32
  3.3.6 Run trained classifier on test dataset .................................................................. 32
  3.3.7 Validate classifier performance ......................................................................... 32
3.4 Classification .............................................................................................................. 34
3.5 Clustering ...................................................................................................................... 34
  3.5.1 Vectorize 3-grams using TF-IDF ............................................................................ 34
  3.5.2 Perform singular value decomposition (SVD) .................................................... 35
  3.5.3 Normalize dataset ............................................................................................... 35
  3.5.4 Perform latent semantic analysis (LSA) ............................................................. 35
  3.5.5 Create k-means object ......................................................................................... 36
  3.5.6 Tweak clustering parameters ............................................................................. 36
  3.5.7 Fit object to transformed data ............................................................................ 36
  3.5.8 Validate clustering performance ....................................................................... 36
3.6 Interpretation ................................................................................................................ 36
  3.6.1 What is the topic configuration of the conversation? ......................................... 37
  3.6.2 What is the DQI category distribution of the conversation? .............................. 37
  3.6.3 What are the important parameters for calculating the DQI category? ............. 37
3.7 Summary ....................................................................................................................... 37
4. Technical implementation details .................................................................................. 38
  4.1 Implementation details ............................................................................................... 38
  4.2 Collection and Processing ......................................................................................... 39
    4.2.1 Download data ..................................................................................................... 39
    4.2.2 Process data ........................................................................................................ 40
    4.2.3 Adding features such as character count and platform .................................... 41
  4.3 Manual Labelling ........................................................................................................ 44
  4.4 Classifier Training .................................................................................................... 47
    4.4.1 Split the labelled dataset .................................................................................... 47
### Chapter 1. Introduction

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.2</td>
<td>Vectorize the words</td>
<td>47</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Create random forests object and tweak classifier parameters</td>
<td>48</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Fit training set</td>
<td>48</td>
</tr>
<tr>
<td>4.4.5</td>
<td>Run trained classifier on test dataset</td>
<td>48</td>
</tr>
<tr>
<td>4.4.6</td>
<td>Validate classifier performance</td>
<td>49</td>
</tr>
<tr>
<td>4.5</td>
<td>Classification</td>
<td>50</td>
</tr>
<tr>
<td>4.6</td>
<td>Clustering</td>
<td>50</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Vectorize 3-grams using TF-IDF</td>
<td>51</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Perform singular value decomposition (SVD)</td>
<td>51</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Normalize dataset</td>
<td>51</td>
</tr>
<tr>
<td>4.6.4</td>
<td>Perform latent semantic analysis (LSA)</td>
<td>51</td>
</tr>
<tr>
<td>4.6.5</td>
<td>Create k-means object</td>
<td>51</td>
</tr>
<tr>
<td>4.7</td>
<td>Algorithm and parameter validation</td>
<td>52</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Classification validation</td>
<td>53</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Clustering validation</td>
<td>53</td>
</tr>
<tr>
<td>4.8</td>
<td>Interpretation</td>
<td>53</td>
</tr>
<tr>
<td>4.8.1</td>
<td>What is the topic configuration of the conversation?</td>
<td>54</td>
</tr>
<tr>
<td>4.8.2</td>
<td>What is the DQI category distribution of the conversation?</td>
<td>55</td>
</tr>
<tr>
<td>4.8.3</td>
<td>What are the important parameters for calculating the DQI category?</td>
<td>56</td>
</tr>
<tr>
<td>4.9</td>
<td>Summary</td>
<td>57</td>
</tr>
</tbody>
</table>

### 5. Design considerations

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Classifier selection</td>
<td>58</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Random Forests classification</td>
<td>59</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Support Vector Machine (SVM) Classification</td>
<td>60</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Logistic Regression Classification</td>
<td>60</td>
</tr>
<tr>
<td>5.1.4</td>
<td>Results of the classifier selection process</td>
<td>60</td>
</tr>
<tr>
<td>5.2</td>
<td>Selection of candidate classification algorithms</td>
<td>61</td>
</tr>
<tr>
<td>5.3</td>
<td>Selection of k-means clustering algorithm</td>
<td>63</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary</td>
<td>63</td>
</tr>
</tbody>
</table>

### 6. Instantiation of framework

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Data collection and processing</td>
<td>65</td>
</tr>
<tr>
<td>6.2</td>
<td>Adding features</td>
<td>70</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Distribution of comment category</td>
<td>70</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction

6.2.2 Distribution of character count ................................................................. 70
6.3 Manual labelling ............................................................................................... 71
  6.3.1 DQI Indicator 1: Participation ...................................................................... 71
  6.3.2 DQI Indicator 2: Level of justification ........................................................ 72
  6.3.3 DQI Indicator 3: Content of justification ..................................................... 73
  6.3.4 DQI Indicator 4: Respect ............................................................................. 73
  6.3.5 DQI Indicator 5: Counterarguments ........................................................... 74
  6.3.6 DQI Indicator 6: Constructive Politics ....................................................... 74
  6.3.7 Classifier training ....................................................................................... 75
6.4 Classification and clustering ............................................................................ 78
  6.4.1 Bernie Sanders dataset analysis ................................................................. 78
  6.4.2 Republican Townhalls dataset analysis ...................................................... 79
  6.4.3 Justin Trudeau Townhalls dataset analysis ................................................ 80
  6.4.4 EULive dataset analysis ............................................................................ 81
  6.4.5 Trudeau Saskatchewan Townhall analysis ................................................ 83
6.5 Node instantiation ............................................................................................ 84
  6.5.1 Random Forests classification using node-level features ............................ 87
6.6 Summary ......................................................................................................... 89

7. Further discussion of results .............................................................................. 91
  7.1 DQI category distribution ............................................................................. 91
  7.2 Topic configuration ....................................................................................... 94
  7.3 Most important classifier features .................................................................. 97
    7.3.1 Character count ....................................................................................... 98
    7.3.2 Platform .................................................................................................. 99
    7.3.3 Political words ....................................................................................... 100
    7.3.4 Commenter frequency .......................................................................... 101
    7.3.5 Eccentricity ......................................................................................... 102
  7.4 Summary ..................................................................................................... 103

8. Conclusion ....................................................................................................... 104
  8.1 Contributions summary ............................................................................... 104
  8.2 Evaluation and limitations ........................................................................... 105
    8.2.1 Evaluation ............................................................................................ 105
    8.2.2 Limitations ............................................................................................ 106
Chapter 1. Introduction

8.3 Further usage................................................................. 106
8.4 Future work................................................................. 106

9. Appendix A: Results from experimental project phase................................. 108
  9.1.1 Random Forest classification using bag-of-words only............................ 110
  9.1.2 Random Forests classification using character count and comment category only.. 111
  9.1.3 SVM classification.................................................................. 111
  9.1.4 Logistic Regression classifier...................................................... 113
  9.1.5 SVM classification using node-level features.................................... 115
  9.1.6 Logistic regression classification using node-level features.................... 117

10. Appendix B: Comment Examples.......................................................... 119

11. Appendix C: DelibAnalysis Codebase......................................................... 120

12. References......................................................................................... 126

Figures

Figure 1: The deliberative process (Fournier-Tombs, 2013) ........................................ 16
Figure 2: An open, undirected network ..................................................................... 19
Figure 3: Social network (left) versus discussion network (right)............................. 20
Figure 4: DelibAnalysis Framework......................................................................... 27
Figure 5: Illustration of precision and recall using low DQI metric ........................... 33
Figure 6: Representation of network structure.......................................................... 41
Figure 7: Example decision-tree.............................................................................. 59
Figure 8: Comment count by category in the training dataset................................ 70
Figure 9: Comment count by number of characters in the training dataset................ 71
Figure 10: Comment count by participation score in the training dataset.................. 72
Figure 11: Comment count by level of justification score in the training dataset........ 72
Figure 12: Comment count by content of justification score in the training dataset...... 73
Figure 13: Comment count by respect score in the training dataset.......................... 74
Figure 14: Comment count by counterarguments score in the training dataset........... 74
Figure 15: Comment count by constructive politics score in the training dataset......... 75
Figure 16: Top 50 features by importance for random forests classifier using hybrid features ...... 77
Figure 17: Discourse quality prediction (random forests) for Bernie Sanders dataset........ 78
Chapter 1. Introduction

Figure 18: Discourse quality prediction (random forests) for Republican Townhall dataset ..........79
Figure 19: Discourse quality prediction (random forests) for Justin Trudeau Townhalls dataset....80
Figure 20: Discourse quality prediction (random forests) for EU Live dataset ..........................82
Figure 21: Discourse quality prediction (random forests) for Trudeau Saskatchewan townhall
          dataset ............................................................................................................................83
Figure 22: Top node-level features by importance .................................................................88
Figure 23: Correlation between discourse quality and topic clusters .....................................97
Figure 24: Character count by DQI score in the training dataset ..............................................99
Figure 25: Character count by discussion platform in the training dataset ..............................100
Figure 26: Comment count by commenter frequency per DQI group .......................................102
Figure 27: Comment count by eccentricity by DQI group ........................................................103
Figure 28: Top 50 features by importance for SVM classifier using hybrid features ...............113
Figure 29: Top 50 features by importance for logistic regression classifier using hybrid features 115
Figure 30: Top features by importance using SVM classifier and node-level features ..............117
Figure 31: Top features by importance using logistic regression classifier and node-level features ..................................................................................................................................................118

Tables

Table 1: Data collection guidelines for using an ethnographic framework ................................ 4
Table 2: Reliability and validity of research ..............................................................................7
Table 3: Definition of deliberation (Gold et al, 2013) .................................................................14
Table 4: Discourse Quality Index (DQI) code book .................................................................17
Table 5: Precision and Recall .....................................................................................................33
Table 6: List of Python packages ............................................................................................38
Table 7: Comments with graph structure ...............................................................................40
Table 8: Adding platform features ..........................................................................................41
Table 9: Adding character features ........................................................................................41
Table 10: Adding node features ..............................................................................................42
Table 11: Manual DQI coding ..................................................................................................45
Table 12: From DQI score to category .....................................................................................46
Table 13: Comment vectorization ..........................................................................................47
Table 42: Classification report using 3 DQI categories, random forests classifier and bag-of-words features .............................................................................................................................................................................................. 110
Table 43: Prediction vs actual comparison using 3 DQI categories, random forests classifier and bag-of-words features .............................................................................................................................................................................................................................................................................................................................................................................................. 110
Table 44: Classification report using 3 DQI categories, random forests classifier and quantitative features .............................................................................................................................................................................................................................................................................................................................................................................................. 111
Table 45: Prediction vs actual comparison using 3 DQI categories, random forests classifier and quantitative features .............................................................................................................................................................................................................................................................................................................................................................................................. 111
Table 46: Classification report using 3 DQI categories, SVM classifier and hybrid features .............................................................................................................................................................................................................................................................................................................................................................................................. 112
Table 47: Predicted vs actual report using 3 DQI categories, SVM classifier and hybrid features .............................................................................................................................................................................................................................................................................................................................................................................................. 112
Table 48: Classification report using 3 DQI categories, logistic regression classifier and hybrid features .............................................................................................................................................................................................................................................................................................................................................................................................. 113
Table 49: Predicted vs actual report using 3 DQI categories, logistic regression classifier and hybrid features .............................................................................................................................................................................................................................................................................................................................................................................................. 114
Table 50: Classification report using SVM classifier and node-level features .............................................................................................................................................................................................................................................................................................................................................................................................. 115
Table 51: Prediction vs actual report using SVM classifier and node-level features .............................................................................................................................................................................................................................................................................................................................................................................................. 116
Table 52: Classification report using logistic regression classifier and node-level features .............................................................................................................................................................................................................................................................................................................................................................................................. 117
Table 53: Predicted vs actual report using logistic regression classifier and node-level features .............................................................................................................................................................................................................................................................................................................................................................................................. 118
1. Introduction

This study begins with the premise that deliberation, the discursive process by which ideas are shaped and formed, enables citizens to converge on given issues, moving from a collection of distinct opinions to a consensual direction. It then follows that engagement must be made with political representatives in order to ensure that the outcomes of these deliberative processes are reflected in legislated action. Public consensus, deliberations and a variety of opinions should be taken into account by political representatives in a healthy democracy. Beyond that, citizens need this two-way communication to receive information that will help them to shape their ideas, and to hold the representatives accountable for their actions.

Although some theorists argue that the only way in which citizens can communicate their opinions is through a vote, most claim that deliberative democracy allows for a much more thorough and healthy exchange of ideas (Chambers, 2009). That is, deliberation and communication allows a society to think, producing more sound and representative legislative outcomes.

But where is this deliberation to take place? In his important book The Structural Transformation of the Public Sphere, Habermas (1961) mentions London and Paris cafés during the late 19th and early 20th centuries as an example of lively and effective fora for public deliberation. Citizens congregated there to hold long and argumentative discussions, often in the presence of politicians and journalists. Habermas argues that during the 20th century, these fora died out, and were replaced with one-way communication technologies such as the radio or the television, through which citizens received information passively.

In the last decade, however, it has become abundantly clear that there is now a new forum for public deliberation, and that is the Internet. In particular, social media platforms such as Twitter, Facebook and YouTube (to name but a few) and blogs (of which there are millions), allow citizens to communicate with each other and with their representatives directly. Although they are not face-to-face, the barrier to social media participation is becoming less and less. Many argue that although the digital divide still exists, social media platforms allow people to communicate and
Chapter 1. Introduction

interact with each other who would never otherwise meet. Case-in-point: Facebook now boasts two billion users worldwide – or more than one quarter of the entire human population.\(^1\)

Political communication and deliberation online can be complex and takes on a variety of forms. What interests us here, however, is deliberation with political representatives. In the pre-digital age, politicians who wished to discuss with or be accountable to their constituents would typically hold periodic in-person Townhalls. Today, although these still occur, it is more common for a politician to post informal comments on their own page and have their constituents respond asynchronously. Online Townhalls do exist, however, when a more formal structure is put together for deliberative purposes. This structure might include formal questions, a moderator, a time-period, a recap of items discussed, for example (Lukensmeyer and Brigham, 2002).

However, the notion of the public has also changed. Discussions online can be targeted to specific interest groups; even to specific individuals. Deliberation thus happens online on a multitude of platforms, with a variety of different groups, in informal, continuous formats and on formal, time-bound ones.

In light of Habermas’ public sphere theory, therefore, it is relevant to ask what is the quality of online deliberation. Is there a way to measure and explore these deliberations to better understand the state of deliberative democracy?

The Discourse Quality Index (DQI), developed in 2003 by Steenbergen, Bächtiger et al for offline parliamentary deliberations, provides concrete indicators based on Habermasian thought. It includes measures for participation, respect, counterarguments and justification, and is a well-established tool for discursive quality analysis. Current practice, however, often consists of manually annotating deliberations rather than using computational methods. In addition, the analysis of users on social media has often captured interactions and depth of discussion separately from content analysis. Overall, current techniques look at content quality, graph characteristics and dynamic behavior in isolation, rather than as components of the same analysis.

This study therefore presents DelibAnalysis, an automated discourse quality measurement framework, which allows us to examine current online discussions and better understand how

citizens deliberate with each other and with their candidates on social media platforms such as Facebook and blogs.

In so doing, this study answers the following questions:

- What factors determine whether an online political discussion is deliberative?
- How can we automatically evaluate the deliberative quality of an online political discussion?
- What other means of understanding deliberative quality might be considered?

In order to answer these questions, the thesis will:

- Present a methodology and tool for the automatic analysis of the Discourse Quality Index (DQI) of social media comments, using supervised machine learning techniques;
- Propose an unsupervised machine learning methodology for understanding the quality of deliberation on social media.
- Implement the above framework to determine the impact of the impact of various variables on online deliberation;

Cases from the United States, Canada, the United Kingdom and the United Nations were chosen for this study due to the high availability and diversity of data that would improve potential outcomes. The main contribution of this study is an analytical tool which will allow for a deeper understanding of deliberation processes, the relationship between discussion quality, content and interactions. This will substantially increase the understanding of the outcomes of political deliberation and the linkages between various levels of participatory quality online. The study also contributes our own findings on deliberative quality online when applying the framework, highlighting the features that make for better or poorer discourse quality in our use case.

### 1.1 Methodology and approach

This study uses a cross-sectional, interpretivist approach, whereby a variety of datasets collected during one time period as a means of ensuring consistency and structure in the results. According to Bryman (2015), this approach is a common alternative to the case study in social science and can be used to detect societal patterns.
Chapter 1. Introduction

While methods used for analysis were quantitative and empirical (the data was analyzed quantitatively by machine learning algorithms) the overall approach was constructivist, in that it relied on a qualitative framework, the DQI, as well as researcher interpretation for both supervised and unsupervised analyses. Additionally, it used an inductive approach - rather than starting with a theory and a hypothesis which would be tested (the deductive approach), it gathered a broad range of observations and findings, which would then be used to feed into political theory (Ritchie et al, 2013). This framed the study as an exploratory, rather than empirical research project (Bryman, 2015).

Although it is far easier to define quantitative than qualitative research (Ritchie et al, 2013), we may start by describing what it is not: a study using solely quantitative, statistical methods for empirical objectives. Many qualitative researchers do, in fact, use statistical methods to analyze the data that they have collected (Bryman, 2015; Luker, 2008; Ritchie et al, 2013). However, a key difference draws in the way that the data is collected, which is often by observation - seeking to obtain human artifacts that will help explain the social world and the meaning that humans associate to its various components.

Observational social data, or passive data collection, as was done for this study, is a research activity that can be categorized as online ethnography. Androutsopoulos (2013) identifies a subset of online ethnographic studies that focus on everyday life on the Internet, which involves only online data collection, as opposed to the use of Internet in everyday life, which would involve offline data collection as well. This certainly applies here, as our study narrows in on a key element of the human experience - democratic participation - as it is experienced online.

Androutsopoulos (2013) identifies six guidelines for data collection online within an ethnographic framework. These are described below, along with the way in which they were addressed in our study.

Table 1: Data collection guidelines for using an ethnographic framework

<table>
<thead>
<tr>
<th>Practice-derived guidelines for systematic observation (Androutsopoulos, 2013)</th>
<th>Application in our study</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>Examine relationships and processes rather than isolated artefacts</strong></th>
<th>Comments were examined as part of discussions, or deliberation processes, and were analyzed as such, particularly when it came to network analysis and clustering analysis.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Move from core to periphery of a field</strong></td>
<td>We started with Facebook Town Halls, which had been at the center of philosophical discussions on online democratic participation since the 2016 US Election, and gradually added less well-known deliberative platforms such as the one produced by UNDP for the Sustainable Development Goals.</td>
</tr>
<tr>
<td><strong>Repeat observation</strong></td>
<td>Where possible, multiple discussions of the same category (Trudeau Townhalls and Republican Townhalls) were obtained.</td>
</tr>
<tr>
<td><strong>Maintain openness</strong></td>
<td>Although some initial hypotheses were held, primarily that not all social network conversations would be of high deliberative quality, an effort was made to allow findings to shape the further direction of the research.</td>
</tr>
<tr>
<td><strong>Use all available technology</strong></td>
<td>This guideline was interpreted to mean “observe all online discussion platforms that are accessible”. This was not possible within the boundaries of our study, however effort was made to obtain data from a cross section of platforms - social media, online chats, and blogs.</td>
</tr>
<tr>
<td><strong>Use observation insights as guidance for further sampling</strong></td>
<td>This was particularly relevant in the case of the UNDP discussion data, which was sampled after the original data collection period. This</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction

was done in order to add node-level features to a new classifier, and also to observe the behavior of the original classifier on a different platform.

The ethnographic approach, which positions societal studies within a cultural framework, is particularly relevant when looking at the broader research question, which is the discourse quality of online deliberation and the ways in which this discourse might contribute to the public sphere.

Finally, one should note that the methodological experiments in machine learning were both an attempt to contribute to the field of discourse analysis research by providing a new analytical tool; and an attempt to understand the public sphere online using a volume of data that would be difficult to analyze using currently existing tools. In this second objective, we were influenced by Weber (Ritchie, 2013), whereby we attempted to answer the research questions using two methods. The first is the direct observational approach, which was served by the supervised, classification exercises. The second is the explanatory or motivational understanding, whereby we used unsupervised learning to extract and interpret meaning from the deliberative actions.

1.1.1 Reliability and validity of research

The research design was evaluated using the concept of reliability and validity, which, although originating from quantitative research (Golafshani 2003), now consist of a widely accepted framework for social science research (Yin, 2003; Bryman, 2015; Golafshani, 2003). Reliability is the extent to which the research in question can be repeated or replicated by other researchers. Validity, in turn, is the way in which the research design actually answers the research questions it intends to answer (Golafshani, 2003). Bryman (2015) further identifies three types of validity - construct validity, internal validity and external validity. Each of these is summarized in turn in the table below.
Chapter 1. Introduction

Table 2: Reliability and validity of research

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
<th>Solution for this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability or replication</td>
<td>Could someone else do the study with the same data and methods and yield the same results?</td>
<td>Use machine-learning algorithms in the analysis and provide open source data and scripts.</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Does the approach really measure what it is supposed to be measuring?</td>
<td>Use a pre-existing measure (the DQI) as a basis for the analysis.</td>
</tr>
<tr>
<td>Internal validity</td>
<td>Does any implication of causal relationships between two variable hold water?</td>
<td>Any causal relationships, for example the weight of a given feature in determining the DQI score, are validated by the algorithms by which they are developed.</td>
</tr>
<tr>
<td>External validity</td>
<td>Do the results of the study apply in a broader, more general context?</td>
<td>One of the objectives of this research is to provide analytical tools that can be used for broader research. The analytical conclusions of the research are examined in the broader context of the public sphere and democratic discourse more generally.</td>
</tr>
</tbody>
</table>

Reliability or replication: Great care was taken in this research to use publicly available data, which could be collected in the same manner by other researchers. Additionally, the data as collected for this research will be made public as an open data corpus, in the case that the platforms that are being used for data extraction remove the raw data from circulation. The scripts written for data manipulation and analysis (as will be described below), are also made public and any software...
package that is used is made public under a creative commons license. Therefore, by using the original data and running the scripts as provided, it will be possible to replicate the exact outputs of all the machine learning exercises.

Construct validity: Because this study used at its core the DQI, which, as we have seen in Chapter 2, is widely regarded as a valid analytical framework in the field, we consider that there is enough consensus on the methodology to give us construct validity.

Internal validity: This was addressed by using already established machine learning software for analysis. All of the packages are part of the Python software modules, which are widely used in academia for computational analysis. The network analysis exercise is done using the Networkx package (Networkx, 2017), which is an extensive library which Hagberg et al (2008) demonstrate is a powerful tool for scientific computation. The machine learning exercises themselves, including both the classifier training and the clustering work are done using Python SciKit (Python SciKit, 2017), also widely used in machine learning research. Any additional external module used in the scripts is listed in the Appendix and has been approved by the Python Programming Language.

External validity: A core tenet of this study was the attempt to create reusable tools which would apply to similar research questions. The other objective was to be able to generalize research results in such a way as to make a contribution to the field of public discourse analysis more broadly.

### 1.2 Contributions

The main contributions of this thesis are twofold. First, it provides an analytical framework, DelibAnalysis, which enables researchers to evaluate the discourse quality of an online political conversation of their choice. Second, in implementing the framework on a dataset of Facebook and blog comments, it extracts the features with the highest contribution to discourse quality and uses them to inform suggestions for the design of platforms that enhance deliberative quality.

### 1.3 Research ethics
According to common research practice (CIHR et al., 1998), although this study involves human data, it only relies on already published, publicly available data. All participants in the online deliberations have done so knowing that their content would be available to the public. Thus, there was no requirement to ask the consent of the participants for this study.

1.4 Current philosophical debates

Social science researchers have argued that being aware of current philosophical debates in a study’s area of inquiry is key to appropriate research design (Ritchie et al., 2013). In this vein, this section discusses recent political events and their impact on news and public discourse related to online deliberation and citizen participation in democracy.

In the last few years, a number of significant political events have triggered world-wide debates on political discourse and political participation. First, the election of the populist businessman Donald Trump, in the United States, surprised many, and was seen by some as a sign that the American democratic system is in decline (Geraghty, 2016).

One of the reasons for this has been a perceived lack of options for voters. That is, the candidates on offer during the election did not reflect the various political ideologies of their electorate (Morgan, 2016). A number of legislative issues in the United States, including job creation, healthcare, immigration and the environment, have since been perceived as showing a disconnect between the will of an important segment of the population and their political representatives (Morgan, 2016). Moreover, while minority interests may appear to have been most discarded (The Economist, 2016), some of these issues, particularly healthcare and taxation, are of concern to a majority whose democratic right to shape the legislative direction of the country appears to be in question.

Another democratic threat has been captured in the popular debate over “fake news”. A critical precursor to democratic deliberation is access to information, which enables citizens to feed in directly to the political process and the issues at hand (Aalberg and Curran, 2012). However, in early 2017, the American President began regularly referring to traditional news outlets as “fake news” - news articles that are verifiably false, and intentionally so (Allcott and Gentzkow, 2017). A Gallup Poll showed a decline in American trust and confidence in mass media in the last few years, with a particularly steep decline among Republicans in 2017 (Allcott and Gentzkow, 2017).
Chapter 1. Introduction

alternative - articles disseminated on blogs or social media, have been shown to contain no fact-checking, editorial judgment or content filtering (Allcott and Gentzkow, 2017).

Ambiguity of information for citizens has not only come from within the United States. In June 2017 the US Federal Bureau of Investigation (FBI) confirmed that the Russian Government had tampered with the US Election - by, amongst other things, disseminating fake news articles on social media.

Meanwhile, other countries faced their own democratic challenges. In 2016, in a controversial referendum, the United Kingdom voted to exit the European Union, a result that, again, was critiqued by many as reflective of the poor state of the democratic system. Finch-Lees (2017) claims that the British vote was manipulated through false information, namely on the responsibilities of the EU and the impact that its membership was having on the people.

At an international level, the United Nations adopted the Sustainable Development Goals (SDG) aimed at increasing the impact of the Millennium Development Goals (MDG) adopted in 2000. The international organization had been criticized, in 2000, of adopting the MDGs unilaterally, without any citizen consultation. The SDGs were therefore adopted with much broader citizen participation, which also significantly increased the number of targets (from 10 to 17), and possibly the difficulty of implementation.

With these challenges to the validity of the flow of information from politicians to citizens, it is unsurprising that difficulties in deliberation - the participation of citizens based on that information and the means by which they formulate and communicate their legislative opinions - would also face some difficulties.

Claudia Chwalisz, author of The People's Verdict, discussed the current state of online deliberation in an interview in July 2017. She presented the current political climate as one that faces an imbalance between the hierarchical, archaic political institutions that are in place, and the modern, open, networked and digital sphere in which their citizens operate (Chwalisz, 2017).

This disconnect has an effect in two key communications channels - it appears that political information no longer flows in a straightforward manner from politicians to citizens; and, in turn, 

---

2 http://time.com/4826603/russia-fake-news-us-election-meddling/
there are not always clear modes for citizens to communicate and deliberate with their politicians, other than voting at the end of each term. Aalberg and Curran (2012) note that it is increasingly easy for citizens to ignore political news. On the Internet, people select both the media they consume and the individuals in their networks, making the late night general public newscast a thing of the past. Oremus (2017) notes that between one and two fifth of the American population regularly obtain their news from Facebook. More disturbingly, Grasseger and Krogerus (2017) detail how social media analytics companies have worked with election candidates to craft bespoke news items for each potential voter on the platform, giving them an extremely narrow (and often misleading) political perspective.

On the other hand, Chambers and Costain (2000) note that, in mass democracies (as opposed to Ancient Greek or city-level democracies) very little public deliberation can really happen face-to-face. Barriers such as travel distance, room capacity and time can make it almost impossible for many citizens to attend public deliberation events. Social media, therefore, has filled an important gap in allowing citizens to discuss and form opinions, both amongst themselves and with their representatives. Nevertheless, when, in February 2017, Republican representatives held in-person and online Town Halls to, among other things, take the pulse of their constituents on an overhaul of the healthcare system, questions were raised as to the means by which the responses and comments would be used to shape legislation³.

It is from an attempt to contribute to the understanding of the politician-citizen relationship in an era of change and confusion that this study was born.

1.5 Structure of the document

In Chapter 1 we provided an introduction to the study, outlined the research questions and discussed the methodology that would be employed in order to answer them. We also provided an overview of the study in societal context and why it is relevant.

In Chapter 2, we will provide a review of related literature, focusing specifically on political discourse and deliberation, the Discourse Quality Index, network analysis, and automated discourse quality analysis.

³ As observed by researcher in dataset introduced in Chapter 5.
In Chapter 3, we will present the DelibAnalysis framework, outlining the steps to be taken in order to conduct an automated discourse quality analysis in an online context.

In Chapter 4, we will outline the technical implementation of the framework, along with sample code and explanations of the full code found in the Annex.

In Chapter 5, we will provide a justification for this framework, listing the steps that were taken to arrive at the methodology and examining alternative machine learning algorithms to the one that was finally selected.

Chapter 6 will describe our implementation of the DelibAnalysis framework described in Chapter 3 and 4. We will use a dataset made up of a broad set of Facebook and blog comments in order to provide insight into a variety of different discussions.

In Chapter 7, we will discuss the results from this implementation, discussing the most important features used in order to obtain discourse quality in our dataset, and what this might mean for the design of a deliberative platform.

Finally, in Chapter 8 we will conclude our study, summarize its contributions and limitations and propose steps for further work.

1.6 Summary

In this chapter, we introduced the problem statement and the research questions for the study. We then moved on to the methodology and approach, the contributions of the study and the research ethics. We also discussed current philosophical debates of relevance to our study and gave a structure of our document. The following chapter will provide a review of literature relevant to our study.
2. State of the Art

This section examines research of relevance to our study. It first examines themes in political theory, particularly political discourse and deliberation, and how these theories have developed over time. It then examines the Discourse Quality Index (DQI), which is used as a basis for formalizing the theories examined in the first section. It then describes the field of network analysis and how network theory can provide additional insight into discussion threads. Finally, it looks at automated discourse quality analysis, building from the DQI section but looking at computational, rather than manual modes of analysis.

2.1 Political Discourse and Deliberation

Deliberation is a communication framework through which participants reach an optimally representative and efficient decision. Plato, in the 4th century BCE, framed the concept as *dialectic*, the means by which citizens of a democracy would rationally examine diverse opinions until a consensus could be reached (Plato, 1992). In classical Greece, deliberation took place in person, and all adult free men in a city-state were allowed to participate. Issues of concern to all citizens, such as war and infrastructure projects could thus be debated openly until a solution emerged. In modern democracies, where states can contain hundreds of millions of citizens, this form of direct democracy is operationally impossible. Thus, two levels of decision-making occur. In the first, citizens vote in order to elect representatives. In the second, the representatives pass legislations, each ideally bearing in mind the interests of their own group of constituents. As Chambers (2003) explains, these divisions highlight two different streams in democratic theory. In the first, citizens are only involved in the political process through their vote. If deliberation takes place, it should only occur at the representative level, in the House of Representatives or the Parliament. In the second, however, citizens also deliberate, in smaller groups, in order to form the ideas that their representative will then push forward into legislation. According to Arendt (1998) and Habermas (1961), deliberation is critical to true democracy and citizenship.

Modern discourse theory is grounded in two of Habermas’ works in particular – *The Structural Transformation of the Public Sphere* (1961) and *The Theory of Communicative Action* (1984). In the first, Habermas describes the role of the public sphere, which is, in large part, to provide a space for
Chapter 3. State of the Art

citizen deliberation to operate. He argues that, in the first half of the 20th century, this sphere has been eroded by the television, the radio and public apathy (1961). In the second, he delves deeper into the analysis of citizen communications and the differentiation between political discourse and other forms of dialogue (1984). As we will see further in the literature review, Habermas’ first work has often been used as a basis for further research on the impact of technological advances on deliberation. His second work has been used to provide insight into the measurement of deliberation quality, which may be used to evaluate the contribution of a deliberative platform to democratic discourse.

Because modern deliberation, particularly at the citizen level, takes place in diverse locations and amongst varied sub-groups, both formally and informally, a definition of terms is important. This definition allows us to identify whether a discussion is a form of deliberation or if it is, in fact, another form of communication.

Fairclough (2013) chooses a broad definition of deliberation, arguing that democracy is about “making choices [...] in response to circumstances and goals, is about choosing policies, and such choices and such actions which follow from them are based upon practical argumentation”. Other authors (Van Dijk, 1997; Davies and Chandler, 2011) also use a relatively informal definition of deliberation, while still others (Chambers, 2003; Gold et al, 2013) argue that it has developed a formal and rather empirical shape over time. Gold et al (2013) compiled 70 or so of these definitions into a table (below), which this study draws upon in its data selection process.

Table 3: Definition of deliberation (Gold et al, 2013)

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition from Gold et al (2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Deliberation is a communicative process that aims at taking a decision (or recommendation) on collectively binding rules or public projects. The substantive goal is to achieve the common good and universality of rules.</td>
</tr>
<tr>
<td>Procedure, institutional properties</td>
<td>As far as possible, the procedure:</td>
</tr>
<tr>
<td></td>
<td>● Should take place in public (or at least be transparent to the public)</td>
</tr>
</tbody>
</table>
### Individual behavior

**Individual participants:**
- Should behave truthfully (authentically)
- Should communicate impartially
- Should behave respectful (sic) towards other persons and their positions, demands, proposals and argument
- Should be open to become convinced by the better argument; that is, to change opinion and to sacrifice individual preferences in favor of the common good

### Communication

The communication should be based on reason, that is it
- Should be based on information as complete as possible about facts, norms, values and preferences of those participating or concerned by the decision
- Should be based on the argumentative justification of all positions and proposals
- Should lead to conclusions only if based on the power of the better argument; a good argument is considered to be empirically correct; logically consistent, and/or to refer to a universal/impartial and valid norm or value

### 2.2 The Discourse Quality Index

The Discourse Quality Index (DQI) was developed in 2003 by Steenbergen et al in order to evaluate parliamentary proceedings in the British Parliament. Although it evaluates representative deliberation rather than citizen deliberation, it has nevertheless provided a useful framework to deliberation analysis, providing a formal measurement methodology based on Habermasian theory. It allowed subsequent researchers to quantify discourse and to measure it against independent variables, such as deliberative platform (online or offline), structure, questions asked, moderation, etc. For example, Dalhberg (2001) argued not only that Habermas' discourse quality framework could be applied to the online sphere, but that a lot more research in this area was needed. Black et
al (2011) published a study on self-governance through Wikipedia, Davy and Kies (2005) on deliberation in Usenet groups, and Monnoyer-Smith and Wojcik (2012) comparing online and offline discourse quality. Many drew from or used revised versions of the DQI, which will be the approach used in this study.

In order to measure Habermas’ ideal deliberation, which would lead to optimal decision making and participation in the public sphere, Steenbergen et al highlighted six indicators – participation, level and content of justification, respect, counterarguments, and constructive politics. The diagram below (Fournier-Tombs, 2013) shows how each indicator serves as a proxy for measuring the deliberative process. It should be noted, however, that the process of deliberation does not always end in convergence of ideas and consensus. As noted by Cohen (2003), deliberation may lead to the identification and delineation of separate areas of disagreement, which would then be addressed by a vote.

![Diagram of the deliberative process](image)

*Figure 1: The deliberative process (Fournier-Tombs, 2013)*

Participation examined the freedom with which citizens are able to contribute to the discussion. Level and content of justification measured the external elements that the citizens brought into their argument to support it. The researchers also examined whether or not citizens included other participants’ counterarguments in their argument – either to
build upon them or to refute them, whether they respected them, and whether their arguments contributed towards a final consensus. In order to implement this, Steenbergen et al used the following codebook, manually analyzing all citizen contributions (in this case, speeches in the British Parliament) and using an external coder to ensure validity.

Table 4: Discourse Quality Index (DQI) code book

<table>
<thead>
<tr>
<th>Category /Code</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>Interruption of a speaker</td>
<td>Participation is possible</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of justification</td>
<td>No justification</td>
<td>Inferior justification</td>
<td>Qualified justification</td>
<td>Sophisticated justification</td>
</tr>
<tr>
<td>Content of justification</td>
<td>Explicit statement concerning group interests</td>
<td>Neutral statement</td>
<td>Explicit statement of the common good in utilitarian terms</td>
<td>Explicit statement of the common good in terms of the difference principle</td>
</tr>
<tr>
<td>Respect</td>
<td>No respect</td>
<td>Implicit respect</td>
<td>Explicit respect</td>
<td></td>
</tr>
<tr>
<td>Counterarguments</td>
<td>Counterarguments ignored</td>
<td>Counterarguments included but degraded</td>
<td>Counterarguments included-neutral</td>
<td>Counterarguments included and valued</td>
</tr>
<tr>
<td>Constructive politics</td>
<td>Positional politics</td>
<td>Alternative proposal</td>
<td>Mediating proposal</td>
<td></td>
</tr>
</tbody>
</table>

One of the primary critics of the DQI is King (2009). He argues that deliberative quality, as conceptualized by Habermas, is an attempt to objectively quantify the manual coders’ subjective beliefs, and as such, is flawed in terms of construct validity. He further claims that although the DQI extracts insightful deliberative elements from a speech, it does not convincingly make a link
between the content of the deliberation and its decision-making outcome. Gold et al (2013) also cite some implementation challenges, due primarily to the labour-intensive manual coding involved.

Nevertheless, it should be noted that Habermas (2006) argues that the impact of deliberative proceedings on legislative outcomes is, in fact, an empirical question, and cites the DQI as evidence, in context, of this relationship.

While the DQI will continue to be improved upon by successions of social scientists, it has nevertheless been serving as groundwork for a growing number of researchers (Chambers, 2003; Jansen and Kies, 2005; Grönlund and Himmelroos, 2003; Kies, 2010; Black et al, 2011; Klein, 2012; Lord et al, 2013; to name a few). It allowed them to quantify discourse and to measure it against independent variables, such as deliberative platform (online or offline), structure, questions asked, moderation, etc. For example, Dalhberg (2004) argued early on that Steenbergen et al’s (2003) methodology could be applied to the online sphere.; Black et al (2011) published a study on self-governance through Wikipedia; Davy and Kies (2005) on deliberation in Usenet groups; Monnoyer-Smith and Wojcik (2012) comparing online and offline discourse quality; Grönlund and Himmelroos (2003) examined the relationship between the design of deliberative discussions online and the deliberative outcome.

2.3 Network Analysis

Networks (also referred to as graphs), at a basic level, involve examining the relationship (edges) between objects (nodes). They have uses in understanding social, technological and biological phenomena, and also in computation, such as calculating optimal distances between objects.
Network theory has been used for quite some time as a framework to better understand communications, social interactions and communities. In social media analysis, it is often used to understand the relationships between people online and the impact of these relationships on their behavior. Before the advent of online social networks, the concept was used notably in conjunction with "social capital" by Bourdieu (1986) to discuss the changing community landscape in the United States (away from churches, for example), and its impact on the individual. From a technological perspective, networks have been a design framework, a means of connecting components together to allow for the flow of, say, electricity or data. Finally, from a biological perspective, networks are used to explain both internal information exchange (as in the case of a neural network), and interconnectivity of discrete entities (as in the case of a tree’s roots, or even an ecosystem) (Newman, 2011).

In the online sphere, it is useful to differentiate between social networks and discussion networks. In social network analysis, each node, or vertex, is a person in the network. Edges are therefore relationships between these persons. In discussion network analysis, each node is a unit of speech, such as a comment. Edges are relationships between these comments, such as comment hierarchy (whether a comment was made in response to another, for example).
Chapter 3. State of the Art

The conceptualization of social relationships as networks has been an extremely productive pursuit for social scientists (Borgatti, 2009). Like Bourdieu, numerous researchers examined the construct of the relationship between people to explain certain behaviours. An early example is Moreno’s 1934 study of the relationship between school girls’ relationships with each other and their likelihood of running away from a boarding school (Moreno, 1934). In the study of discussion networks, an area of interest is the core discussion network, which examines the relations that people turn to when discussing issues of importance. In such discussions, several authors (Smalls, 2013; McPherson et al, 2006), have found that these discussions tend to involve participants with weak ties, potentially because those outside the person’s core network of family and friends are more likely to provide alternative viewpoints. Conversely, Marsden (1987) had noted a striking lack of heterogeneity in core discussion networks among Americans in the 1985 US Census.

Another way, however, to think about network analysis is to conceive of the discussion, rather than the external relationship between actors, as the network. This is often done when there is no apparent relationship between participants in a discussion, or when the object of analysis is not the overall behavior of an actor but rather one comment or action in particular. Gonzalez-Bailon (2010), for example, compared the political discussion networks in Slashdot forums to discussions of different topics.

In the online world, these are also commonly referred to as discussion threads. Hewitt (2003) writes that discussions by students in online education settings tend to have long threads, showing a bias by the students towards newer posts. When examining the structure of discussion threads, thread depth is often used to measure participant interest in given posts (Gonzalez-Bailon, 2010).
Other areas include roles taken by various participants, such as moderators, influencers, etc. (Katona et al, 2011); response time (Fournier-Tombs, 2013); content and character or word count (Blumenstock, 2008).

Katona et al (2011), noted that influencers in online social networks (those whose comments were most likely to be interacted with), showed an acceleration in adoption rates. That is, the rate of growth of their influence increased over time, as other users seemed more likely to trust content that was already accepted by their peers.

Dubois and Dutton (2012) examined the impact of network structure and nodal behavior on the outcome of Bill C-30, which aimed to protect children from online predators. They found that activity centered around three influencer accounts had a significant impact on “killing” the bill, which was deemed too intrusive.

However, network structures can be examined in different ways, to different results. Dubois and Gaffney (2014), for example, when examining the political tweets of Canadians, found that when measuring network-wide positioning, traditional influencers such as media and politicians were found to be more influential. However, when examining localized interactions in smaller clusters, bloggers and political commentators gained in influence.

Network analysis of online discussions has also drawn from Habermasian theory, in that he argues that the intensity of engagement and volume of participants can lead to stronger deliberative outcomes. For example, in her analysis of the Slashdot forum, Gonzalez-Bailon (2010) found that political discussion networks showed deeper and wider network structures than other types of discussions. Using structural cues for the analysis of networks in social media is quite common, with tie strengths examined by Eveland and Kleinman (2013), amongst others. Analysis of social media networks has become common both in the academic and in the data visualization fields, with the objective of finding dependency factors between network structures and other independent variables (Myers et al, 2014, Suh et al, 2010, Kivran-Swaime et al, 2011, Zaman et al, 2011). In addition to deliberative quality, questions asked included: what makes a particular post more popular (Suh et al, 2010); why do certain individuals gather clustered communities around them (Smith et al, 2014); and what factors lead to the breaking of social ties (Kivran-Swaime et al, 2011).
Quantitative methods for social network analysis draw from the computer science concept of applied graph theory, which allow for the representation and the understanding of the connections between entities in the network. Measuring these connections, their distributions and their segmentation in a given network has allowed researchers to explore correlation or causal linkages between different attributes of the network. Hanneman and Riddle (2005) describes the most common metrics for social network analysis, which include density (how connected are the vertices (Gonzalez-Bailon, 2010), centrality (the importance of a vertex (Smith et al, 2014) and structural holes (are there pairs that are not connected to any others in the network). As seen above (Eveland and Kleinman, 2013), tie strength -- the intensity or reciprocity of a relationship, is also commonly examined.

Re-post cascades, graphs of re-posts or forwards constructed from original comments (Lerman et al. 2012) are an additional useful indicator of an interest sparked by a topic. Research in this field concentrates on predicting single cascade evolutions (e.g. growth and structure) (Cheng et al. 2014), understanding the relationship between cascades and the dynamic changes of the underlying social graphs (graphs of followers) (Myers and Lescovec, 2014), or analyzing cascades taking into account the underlying social graphs (Lerman et al. 2013). Recently, alternative diffusion models based on multi-agent systems are emerging for understanding social influence (Ikeda et al., 2014) or for analyzing urgent diffusion (Rand et al. 2015). A few studies consider the dynamic evolution over time of the social graphs characteristics (Rossi et al. 2013).

Another important research direction in social graph analysis aims at detecting communities (Papadopoulos et al., 2011). Communities are defined relative to a modularity criterion (Newman and Girvan, 2004) that expresses the fact that members of a community must have more relationships with other community members than with non-members. Multiple algorithms, e.g. (Blondel et al., 2008) have been developed to efficiently find communities in very large graphs. Researchers have also studied the meaning of the obtained communities, by comparing them with real-world communities.

2.4 Automated Discourse Quality Analysis

The automation of discourse quality analysis has posed a challenge for researchers. As we have seen, Steenbergen et al (2003) manually coded all of the speeches in their original research.
Approaches to automating this process involved quantifying elements in the text, such as syntactic or semantic elements (commonly referred to as natural language processing), length or node positioning, and then using these as indicators (or features) in statistical or machine learning computations. These computations primarily involved linear regression (a statistical method), classification (a supervised machine learning method) or clustering (an unsupervised machine learning method).

Machine learning methods are commonly used in the social sciences, particularly when it comes to content analysis. In discourse analysis, it allows for the automated annotation of phrases or speech-acts by classifying them based on their features, as well as the clustering of data together based on similarity. Classification exercises fall under supervised learning, which is a class of machine learning algorithms that require an annotated dataset on which to train. They then extract and weight features from the training data to predict the classification of non-annotated data. Clustering algorithms, on the other hand, are used on data that is not previously annotated. This class of algorithms fall under unsupervised learning, and these algorithms will seek to group data together that have similar features and unearth patterns in the dataset (Murphy, 2012).

Lin et al (2009) argued that in online learning, obtaining information on the nature of a comment - whether it is a question, a clarification, and assertion, etc. - would ease the work of the teacher monitoring them. The group used cascade classification, a machine learning algorithm that uses the output of each stage to refine the classifier.

Blumenstock (2008), when faced with the question of evaluating the quality of Wikipedia articles, proposed a very simple metric, word count, finding a significant correlation between the two, where the higher the word count, the higher the quality of the article. As a classification feature, the researchers found that the metric performed well independently of classification algorithm or parameters.

As in network analysis, some of the most detailed work in automated discourse quality analysis has been done in online learning, particularly online grading. ETS, the entity administering university entry tests such as the SAT, GRE and GMAT, has used automated evaluation in its grading of essay answers since 1999 (Burstein et al, 2001). The automated grader uses three categories of features in its classifier - syntactic variety (as determined by part-of-speech tagging), organization of ideas...
Chapter 3. State of the Art

(as identified by a series of words, parts of speech and their placement in the text), and vocabulary usage. This classifier had, according to the authors, shown 97% accuracy in 750,000 grades allocated. An updated version of the tool (Attali and Burstein, 2004) added essay length among the improved feature set. Additional features included lexical complexity (the ratio of words to tokens) and specific vocabulary use.

One of the challenging aspects of discourse quality analysis, as opposed to discourse analysis, is that it is evaluative, rather than merely descriptive. Therefore, although descriptive techniques such as sentiment analysis (Malouf, 2008) have been used extensively to analyze online content, evaluative techniques such as the ones used for exam grading are rarer. discourse quality analysis of other formal speech. Nevertheless, Gifu and Cristea (2013) created an automated analyzer for Romanian political discourse, found that using semiotic natural language processing techniques could prove of value for discourse analysis, particularly when comparing speech patterns of different politicians. Their work drew largely on Pennebaker et al’s (2001) Linguistic Inquiry and Word Count (LIWC) tool, which provides summary variables for input text of various categories. For example, it categorizes professional or scientific writing in terms of analytic, clout, authenticity and emotional tone. While higher scores for each are better, it also provides an average score for writing of that nature for comparative purposes.

Finally, several authors have attempted to measure the quality of online political discourse, which is our object of inquiry. Some, returning to a more formal definition of deliberative quality, have evaluated the outcome of the deliberation, be it increased knowledge on the part of participants, or clear progression from deliberation to decision (Carman et al, 2015). In this construct, a discussion of high deliberative quality is one that has clear positive outcomes. Others, however, noting the importance of measuring deliberative quality, particularly for comparative purposes (one platform versus another, for example), have begun to develop methods to reduce the manual component of discourse quality analysis (Gold et al, 2013; Muhlberger and Stromer-Galley, 2009).

Gold et al (2013), in their study on automating the DQI (Steenbergen et al, 2003), argued that a fully automated process can only be realized on large datasets. They also claimed that the DQI, relying as it does on a highly subjective, manual coding process, presented some challenges when it came to automation. They built a content analysis framework for online discourse, splitting texts into lexical episodes (or phrases) which can then be grouped by meaning. In 2017, they resolved this issue by
developing separate annotation techniques for each of four categories: participation, respect, argumentation and justification, and persuasiveness. Participation was evaluated by calculating the share of turns for each participant in each topic that was discussed. Respect was evaluated by noting the absence of markers of disrespect and impoliteness. In turn, argumentation and justification were evaluated by calculating the presence of causal connectors (for, because, as, since, therefore, etc.) and discourse particles (so, well, you know, I mean, etc.); and persuasiveness by the presence of performative verbs (accept, threaten) or attitude verbs (believe, assume, know).

2.5 Conclusions from state of the art section

As we have seen above, a variety of interesting and detailed work has been done in the field of online discourse quality. However, where we see two gaps: the first is in the development of a more robust machine learning method for the classification of online deliberations. The second is in exploring unsupervised methodologies – ones that would not require a manually coded training set – that would indicate discourse quality.

2.6 Summary

In this chapter, we reviewed the literature of relevance to our study. We began by describing political discourse and deliberation theory (2.1), then narrowed in on the Discourse Quality Index (2.2), then discussed network analysis (2.3) and automated discourse quality analysis (2.4). In the following chapter, we will introduce DelibAnalysis, which is an attempt to fill the analytical gaps present in the literature.
3. DelibAnalysis Framework

In this chapter, we present the proposed framework for analyzing discourse quality using machine learning methodologies. This framework addresses the gaps highlighted in the previous sections, namely an automated means of analyzing discourse quality in both a quantitative and a qualitative way. The framework was developed after comparing several relevant methodologies; a more complete detail of the experimental process is described in Chapter 5.

The framework is meant to be a step-by-step guide to analyzing discourse quality of political discussions found online. It is flexible enough to allow for the use of any platform (blog, social media, chat, etc.), and any language, although the implementation is shown in English. Although it does require the manual annotation of a training dataset, as we will describe below, once the classifier is trained, it can be applied on datasets of any size.

The DelibAnalysis procedure is explained in the swim lane diagram below. It involves six categories – collection and processing, manual labelling, classifier training, classification, clustering and interpretation; each of which will be explained in further detail below. The technical implementation, including sample Python code, will be covered in Chapter 4.
Chapter 3. DelibAnalysis Framework

3.1 Collection and Processing

This category covers three steps: data download, processing and feature selection. The expected output is a single two-dimensional table for analysis.

3.1.1 Download data

Data used for this analysis should be related to political discourse. As we have seen in Section 2.1, political discourse is a specific mode of conversation with political objectives and a certain format, namely, that it is public. In our implementation of DelibAnalysis, which we will discuss in Chapter 5, we selected as political discourse online discussions that took place with elected
representatives, as well as discussions that took the format of public consultations on political topics (in our case, poverty reduction).

Once an appropriate dataset has been selected, it can be downloaded locally. The methodology for this is discussed in the implementation section in Chapter 4. Many public data sources, particularly hosted on social media platforms such as Facebook, Twitter and Reddit, have Application Programming Interfaces (APIs), which allow for the download of publicly available discussions. In the case that an API is not available, it is often necessary to write a script to scrape the data.

3.1.2 Process data

Once the data has been downloaded, it should be pre-processed in order to ease analysis. In DelibAnalysis, we will use a CSV-formatted dataset that is free of non-ASCII characters. The format of the dataset, including column headers, will be further discussed in Chapter 4.

3.1.3 Adding features such as character count and platform

As we will see in the next steps, a significant portion of the DelibAnalysis approach consists of comment analysis, using term-frequency-inverse-document-frequency (TF-IDF) and n-gram analysis. These “word” and “phrase” features will be created automatically using a vectorizing script. However, there is value in other, quantitative features as well. Notably, if data from multiple platforms is used for this analysis, they should be added as features, as well as character count categories. In certain platforms, other features, such as number of likes or number of shares can also be added.

Node-level features can also be added as quantitative features. These imply a graph structure of the comments, in which each comment is a node within a conversation’s architecture. For a more detailed explanation of graphs and networks, please see Section 2.2. Relevant node-level features include degree centrality, betweenness centrality, whether or not a comment is a leaf, commenter frequency and eccentricity. These features can be useful when examining a discussion with a deep hierarchy with multiple embedded responses. Possibly, for example, comments which are responding to the original post may have a different DQI category than comments that are
responding to other commenters, or which themselves have many responses. Each suggested metric is defined in further detail below.

a. **Degree centrality**: A measure of how many neighbors a node has. This is a centrality measure that will, in our case, calculate how many comments directly refer to a comment.

b. **Betweenness centrality**: The number of shortest paths that pass through a given vertex. Slightly different than degree centrality, betweenness centrality is also a centrality measure that ranks how interconnected a node is.

c. **Is leaf**: True if a node has no children. In a graph structure, a leaf is a post to which no one has responded. This measure therefore evaluates whether a post has a response or not, and whether that has any impact on the discourse quality.

d. **Commenter frequency**: number of commenter posts divided by total number of posts. This straightforward measure evaluates how often a commenter has contributed to a given conversation. It may be that commenters who contribute a lot are leaders or informal moderators of the discussion, for example, or conversely, they could also be spamming the conversation.

e. **Eccentricity**: maximum distance between a node and any other node. Like centrality, this measure evaluates the position of the comment in the structure. The more eccentric a comment is, the more likely it is to be a less prominent response.

### 3.2 Manual Labelling

This process involves manually adding DQI labels to a portion of the dataset which will be used to train the classifier. As we have seen, this step is directly inspired by Steenbergen et al’s work (2003) manually annotating speeches in the British Parliament. In our method, we use the indicators provided by Steenbergen et al to arrive at a DQI score, and then extract from that a DQI category. Our classifier will later reverse-engineer the DQI category, by selecting quantitative indicators (features) that are different from the qualitative ones used in manual classification. The steps for manual labelling are as follows.
3.2.1 Selecting a portion of the dataset for manual labelling

In this step, a portion of the dataset is set aside for the manual labelling step. In machine learning, there is no rule of thumb for how big the training dataset should be. If a discussion contains 1 million comments, for example, it may not be scalable to manually label an arbitrary 5% or 10%. However, we are using a random forests classifier, which has been shown to be quite effective on small datasets. What is important, in this case, is to have a representative sample of the dataset. For example, if the dataset contains 50% blog posts and 50% Twitter posts, the training dataset should not just have Twitter posts, but rather a similar ratio of both. A non-scientific range of training set size, depending on the number of commenters, could be 500-5,000 comments. As we will see once we evaluate the classifier, is that if its performance is too low, the training set may have to be increased in size. We therefore suggest to start with a smaller training set and increase it afterwards if necessary.

3.2.2 Adding the score for each of the six DQI indicators

As we have seen, the six DQI indicators are: participation, level of justification, content of justification, respect, counterarguments and constructive politics. The coding manual used by Steenbergen et al is provided in Section 4.4. Each comment should be given a score for each of the indicators, ranging from 0 to 3, depending on the indicator. These indicators are a method for a human evaluator to quantify the quality of discourse, while using subjective, qualitative metrics. The machine evaluator will use very different metrics.

3.2.3 Combining indicator scores to obtain the DQI score

The total DQI score ranges from 0 (worst quality) to 14 (highest quality). Each indicator should be summed to obtain the total, index score.

3.2.4 Converting the DQI score into a DQI category

Once the index score is calculated, it should be converted into a DQI category of low (score is 0-5), medium (score is 6-10) or high (score is 11-14). Classification will perform much better with 3 possible outcomes, rather than 14. If we had kept the index scores, a machine score of 13 when a manual score was 12 would have been evaluated as wrong. In this case, both scores would fall
in the high category and would be evaluated as correct. Understanding the general trend of the score is more useful to us than having the exact score.

3.3 Classifier Training

This process involves training the classifier until the F1 score is satisfactory. Once trained, the classifier will then be used on unlabelled data to automatically obtain the DQI category for each comment.

3.3.1 Split the labelled dataset into training and testing sets

A best practice for dataset splitting is 80% of the dataset for training and 20% for testing. This means that, given a labelled dataset with 1000 comments, 800 should be used for classifier training and 200 should be used for testing. The testing dataset will allow us to evaluate the accuracy of the classifier in predicting the DQI category.

3.3.2 Vectorize words in data

Because the dataset largely consists of a comment, the words or phrases in the comment need to be vectorized. This means that they are converted to a matrix structure in order to be able to quantify the words appearing in a given comment and how important each word is in determining discourse quality. Sample code for the vectorization, along with all the steps for training, is provided in Chapter 4.

3.3.3 Create random forests object

As we will see in Chapter 5, the random forests algorithm was selected because of its good performance with a relatively small dataset (approx. 700 records), compared to two other common classification algorithms – Support Vector Machines (SVM) and logistic regression. A random forests classifier is a set of decision trees (commonly 100), which give a classification for an input data based on a series of randomly selected features. The classification that the forest selects for the input (in our case low, medium or high DQI), is the one that received the most votes from the individual decision trees (Berkeley, 2017). A more complete description of the algorithm is provided in section 5.4.
3.3.4 Tweak classifier parameters

A set of default classifier parameters are provided with the DelibAnalysis code, and can be used as a starting point for any new implementation. However, depending on the particularities of the dataset, some tweaks may lead to better results. A full description of each parameter available in the Python SciKit Learn library is available here: [http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html).

3.3.5 Fit training set to random forests object

Once the classifier object is created, the next step is to train it using the portion of the data set aside as a training set. This step will generate a weighted ranking of features, which will then be used for selecting the classification of unlabelled data. For example, character count may be an important factor in DQI category (the more characters the higher the quality), but the presence of the words “party” and “country” may be less important.

3.3.6 Run trained classifier on test dataset

Once the classifier is trained, it is run against the test dataset, in order to evaluate whether the expected classification corresponds to the actual classification of each comment.

3.3.7 Validate classifier performance

This is then validated using a confusion table, which contains the following statistics: precision, recall and F1. These statistics are measured as follows, giving the example of whether or not a comment has a low DQI (the same exercise would take place as to whether or not a comment has a medium DQI and a high DQI).

- **True Positives (TP)**: Comments that are classified as having a low DQI and actually have a low DQI.
- **True Negatives (TN)**: Comments that are classified as not having a low DQI and actually do not have a low DQI.
- **False Positives (FP)**: Comments that are classified as having a low DQI but actually do not have a low DQI.
- **False Negatives**: Comments that are not classified as having a low DQI but actually have a low DQI.
- **Precision**: $\frac{TP}{TP+FP}$ – The percentage of elements that the classifier selected as having a low DQI that actually have a low DQI.
- **Recall**: $\frac{TP}{FN+TP}$ – The percentage of elements that the classifier selected as having a low DQI compared to all elements having a low DQI.
- **F1 score**: The harmonic mean of precision and recall: $2\left(\frac{1}{1/\text{Recall}}+\frac{1}{\text{Precision}}\right)$

*Figure 5: Illustration of precision and recall using low DQI metric*

In the example illustrating in the figure above, the statistics would be calculated as follows:

*Table 5: Precision and Recall*

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.44</td>
</tr>
<tr>
<td>Recall</td>
<td>0.4</td>
</tr>
<tr>
<td>F1</td>
<td>0.42</td>
</tr>
</tbody>
</table>
In the validation step, if the F1 score is not showing adequate accuracy, the classifier should be recalibrated in order to improve the score. We recommend a minimum of a 0.8 F1 score in order to move on to the next step. If the score is lower, then two options are suggested. First, revisit the classification parameters to ensure that they are appropriate for the dataset. If this still does not improve the results, then it becomes necessary to increase the size of the training dataset. If the size of the dataset is closer to 500 comments, we recommend increasing the size of the dataset until 5000 comments is reached. If at that stage, the F1 score is still not satisfactory, it may be that the training set does not appropriately represent the data in the rest of the dataset. We recommend that the training data matches the rest of the data as much as possible in distribution of character count, platform selection, language(s), and topic range.

### 3.4 Classification

This category covers one step only: running the trained classifier on the unlabelled portion of the dataset. Once the classifier reaches a satisfactory accuracy, it can be run on the rest of the dataset, which is not labelled with the DQI score. This will return the DQI categories for each comment in the dataset, along with the most important feature weightings, which we will analyze further in the following sections.

### 3.5 Clustering

In this process, we obtain a set of topic clusters from the conversation under study. The steps involved are:

#### 3.5.1 Vectorize 3-grams using TF-IDF

This exercise uses an unsupervised thematic analysis, using k-means classification. Similarly to the vectorization step in the classification process, in this step we also vectorize the comments but using 2 or 3-grams, which are phrases of 2 or 3 words each. For example, in a given comment: “I think the middle class should pay less tuition.”, The following 3-grams are extracted: “I think the”, “think the middle”, “the middle class”, “middle class should”, “class should pay”, “should pay less”, “pay less tuition”. This was done in order to obtain themes for each cluster that were more fleshed out and easier to interpret.
K-means therefore requires the data to have features - in this case we used TF-IDF. This measure is the product of two measures - term frequency and inverse document frequency. Term frequency is defined as the number of times that term $t$ occurs in a document, $d$. Inverse document frequency is defined as the rarity of a term in all the documents in the dataset. For example, if the term occurs in each comment, the idf score is zero (Bernadi, 2017). These TF-IDF-weighted n-grams are then used to form quantitative features in sparse matrix representation.

The labelled data is therefore analyzed using k-means as the clustering algorithm and TF-IDF as the feature strategy. In this classification model, groupings, instead of being predicted based on previous labels, are derived from the content of the text (or, in the case of numeric data, the value). Using term-frequency inverse-document-frequency (TF-IDF), we therefore cluster the data by most relevant phrases. The parameters selected for each method are listed in Appendix A.

3.5.2 Perform singular value decomposition (SVD)

SVD is a matrix factorization procedure commonly used as a pre-processing technique for clustering (Wall et al, 2002). The technique allows for a compression of the sparse matrix created in the previous step in order to more efficiently create the clusters.

3.5.3 Normalize dataset

As above, normalization is a well understood technique for clustering pre-processing, particularly for k-means analysis (Virmani et al, 2015). This step models the clusters efficiently to reduce redundancies and diminish the likelihood of irrelevant clusters.

3.5.4 Perform latent semantic analysis (LSA)

LSA is a dimension reduction technique with the objective of assigning topics to each comment, increasing the likelihood that comments with overlapping topics will be assigned to the correct clusters. It is often used with the above steps to ensure the quality of the clusters (Wall et al, 2002).
Chapter 3. DelibAnalysis Framework

3.5.5 Create k-means object

On the pre-processing is complete, we create the k-means object. K-means is a clustering technique widely used in text analysis, which will be further discussed in section 5.3. The objective of creating the k-means object will be to create topic clusters for the discussion based on the 3-gram phrases created in the earlier steps.

3.5.6 Tweak clustering parameters

As in the classification step, default clustering parameters are provided with DelibAnalysis. Namely, we propose 5 clusters per discussion as a baseline. Longer, more complex conversations may require more clusters, and this can be tweaked in the original code provided. All k-means clustering parameters are described in more complete detail here: http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

3.5.7 Fit object to transformed data

In this stage, we now run the clustering algorithm on the transformed data. The output will be 5 clusters (if this hasn’t been changed). In our implementation, we show the top 10 3-gram phrases for each cluster to show how the documents were categorized into each group.

3.5.8 Validate clustering performance

In order to validate the quality of the cluster, we use the silhouette coefficient metric. This metric allows for the validation of the cluster on unlabelled data in particular, to assess whether the clusters are internally consistent. Our implementation provides a score between -1 and 1, where negative values mean that the comments have been assigned to the wrong cluster, 0 means overlapping clusters and 1 is the most distinct and internally coherent clusters.

3.6 Interpretation

In this section, we examine the results of our DelibAnalysis and extract relevant insights. In the DelibAnalysis codebase, we provide means for deriving visualizations for each of the questions. The key section are as follows:
3.6.1 What is the topic configuration of the conversation?

This step provides the top 10 2 and 3-grams for each of the topic clusters in the conversation. This allows for an oversight of the main themes discussed, and a subjective assessment of the political relevance of the conversation. Additionally, a distribution of the number of documents for each cluster is shown, in order to understand how distinct the clusters were and whether there were some topics that were discussed more intensively than others.

3.6.2 What is the DQI category distribution of the conversation?

Simply put, this is a histogram with basic statistics showing the distribution of the DQI categories across all comments in the conversation. This allows us to have an understanding of the quality of the conversation as a whole.

3.6.3 What are the important parameters for calculating the DQI category?

This analysis gives us the “why” as to the DQI scores obtained by the previous step. It gives us a list of parameters by importance, such as eccentricity, character count, and certain phrases, in order to understand why comments would rank as they did. This in particular allows us to further understand what are the elements of an online discussion that would lead to a higher discourse quality.

The following chapter will provide the technical details of the implementation, along with sample code in Python. The complete codebase is available in the Appendix, as well as on GitHub.

3.7 Summary

In this chapter, we introduced the DelibAnalysis framework, which uses classification and clustering techniques to provide a comprehensive discourse quality analysis of a given dataset. We began with a swim lane flow chart that provided an overview of each step in the framework, and followed with a more detailed description. The following chapter will go through these steps again, this time providing technical implementation details using Python and the Scikit Learn machine learning framework.
4. Technical implementation details

In the following chapter, we provide technical implementation details for DelibAnalysis. The implementation was conducted using the Python language and modules which will be detailed below. All code is provided in iPython Notebooks on GitHub (https://github.com/eleonoreft/DelibAnalysis), as well as in the Appendix.

4.1 Implementation details

In order to implement DelibAnalysis, Python 3 should be installed locally, along with the following Python packages:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>nltk</td>
<td>Natural Language Toolkit – natural language processing</td>
<td><a href="http://www.nltk.org/">http://www.nltk.org/</a></td>
</tr>
<tr>
<td>pandas</td>
<td>Pandas – data analysis</td>
<td><a href="http://pandas.pydata.org/">http://pandas.pydata.org/</a></td>
</tr>
<tr>
<td>numpy</td>
<td>Numpy – numeric analysis</td>
<td><a href="http://www.numpy.org/">http://www.numpy.org/</a></td>
</tr>
<tr>
<td>matplotlib</td>
<td>Matplotlib - plotting</td>
<td><a href="https://matplotlib.org/">https://matplotlib.org/</a></td>
</tr>
<tr>
<td>networkx</td>
<td>Networkx – network analysis</td>
<td><a href="https://networkx.github.io/">https://networkx.github.io/</a></td>
</tr>
</tbody>
</table>

A recommended installation for researchers who do not yet have Python installed locally is Anaconda, which at the time of writing contains all of the above modules except for Networkx.
Chapter 4. Technical implementation details

For the purposes of this analysis, once the data is downloaded, it is stored in CSV files locally. For larger datasets, a database could be used. While this research is database agnostic, it is common practice for social media analysis to use a NoSQL database such as ElasticSearch or MongoDB.

### 4.2 Collection and Processing

#### 4.2.1 Download data

Facebook data can be downloaded via the Facebook Graph API. The steps are as follows:

1. Make a request to the API, which will allow for the download in JSON of the first 500 comments:

   ```python
   def make_request(url=test_url):
       r = requests.get(url)
       return r.json()
   
   def download_data(url=test_url):
       r = requests.get(url)
       return r.json()
   
   1. Use the Facebook Graph API to continue downloading data in batches of 500 until there is nothing left:

   ```python
   while req['paging']['next'] is not None:
       new_url = req['paging']['next']
       req = make_request(new_url)
       counter += 1
       with open(args.g+'-'+'str(counter)+'.json', 'w') as fp:
           json.dump(req, fp)
   
   2. Convert the JSON files to CSV. The JSON files can be kept for archival purposes as raw data. The following fields should be used in the CSV file: ‘comment’, ‘commentid’, ‘created_time’, ‘username’, ‘userid’, ‘parent_id’.

   Blog data comes in a variety of formats. Some blogs have APIs that can be called in a way similar to Facebook. Others can be scraped manually or using a custom script.
### 4.2.2 Process data

For most of the project, the only modification that should be made to the data is to clean out non-ASCII characters and emojis. However, the network analysis phase requires a graph structure to the data, as follows:

**Table 7: Comments with graph structure**

<table>
<thead>
<tr>
<th>comment</th>
<th>commentid</th>
<th>created_time</th>
<th>userid</th>
<th>username</th>
<th>parent_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take back the voting restrictions</td>
<td>111111111111111111 11</td>
<td>2017-01-13T15:39:19+0000</td>
<td>101555330742</td>
<td>Gabby Gallant</td>
<td>1111111111 1111111111 1</td>
</tr>
<tr>
<td>Please cut down taxes for middle class</td>
<td>222222222222222222 22</td>
<td>2017-01-12T23:30:07+0000</td>
<td>101526536370</td>
<td>Rana M Imran</td>
<td>1111111111 1111111111 1</td>
</tr>
<tr>
<td>Everything costs too much</td>
<td>333333333333333333 33</td>
<td>2017-01-12T23:31:03+0000</td>
<td>746660392038</td>
<td>Travis Brown</td>
<td>1111111111 1111111111 1</td>
</tr>
<tr>
<td>How can we be sure the vote count isn't rigged?</td>
<td>444444444444444444 44</td>
<td>2017-01-12T23:45:05+0000</td>
<td>235551231765</td>
<td>Becky Turley</td>
<td>22222222 22222222 22222222 2</td>
</tr>
</tbody>
</table>
The table above would yield a network structure as follows. In the following section, we discuss the implementation of node-level features.

![Network Structure Diagram](image)

**Figure 6: Representation of network structure**

### 4.2.3 Adding features such as character count and platform

**Table 8: Adding platform features**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Facebook</th>
<th>Blog</th>
<th>Town Hall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Comment 2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The same should be done with character count. We propose separating the character counts into categories. In our implementation, we create these features as follows.

**Table 9: Adding character features**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Less than 30 chars</th>
<th>Between 30 and 100 chars</th>
<th>Between 100 and 300 chars</th>
<th>Between 300 and 800 chars</th>
<th>Between 800 and 1500 chars</th>
<th>Between 1500 and 3000 chars</th>
<th>More than 3000 chars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Comment 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
There may be cases in which using network analysis to create features may be appropriate for the dataset. For example, a dataset in which there is a significant amount of nested conversations may have varying quality depending on the position of the node or the frequency of the commentor.

The features that will be added to each comment will be as follows:

**Table 10: Adding node features**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Sample value for analyzed node</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree centrality</td>
<td>0.786982</td>
<td>Double</td>
<td>A measure of how many neighbors a node has.</td>
</tr>
<tr>
<td>Betweenness centrality</td>
<td>0.994505</td>
<td>Double</td>
<td>The number of shortest paths that pass through a given vertex.</td>
</tr>
<tr>
<td>Is Leaf</td>
<td>True</td>
<td>Boolean</td>
<td>True if a node has no children.</td>
</tr>
<tr>
<td>Commenter frequency</td>
<td>0.005882</td>
<td>Double</td>
<td>Number of commenter posts divided by total number of posts</td>
</tr>
<tr>
<td>Eccentricity</td>
<td>3</td>
<td>Double</td>
<td>Maximum distance between a node and any other node.</td>
</tr>
</tbody>
</table>

In the first step, we create a graph object from the dataset.

```python
G = nx.Graph()
```

We then append the "Is leaf" feature to each comment. The code below sets each node as "is_leaf = True" and then changes it to "False" if it has a parent node.

```python
for index, row in df.iterrows():
    G.add_node(row['node_id'], attr_dict={'is_leaf':True, 'level':0})
```
Chapter 4. Technical implementation details

G.add_edge(row['node_id'], row['parent_node_id'])

for n in G.edges():
    G.node[n[0]]['is_leaf'] = False

We then create a dictionary for the 'is leaf' feature.

is_leaf = {}
key = 0
for n in G.nodes():
    is_leaf[key] = G.node[n]['is_leaf']
    key += 1

A similar dictionary is created for each of degree centrality, betweenness centrality and eccentricity, using the built in library functions.

degree_centrality = {}
for k, v in nx.degree_centrality(G).iteritems():
    degree_centrality[k-1] = v # Compute the degree centrality for nodes

betweenness_centrality = {}
for k, v in nx.betweenness_centrality(G).iteritems():
    betweenness_centrality[k-1] = v # Compute the betweenness centrality for nodes

eccentricity = {}
for k, v in nx.eccentricity(G).iteritems():
    eccentricity[k-1] = v

Finally, a dictionary is created for commentor frequency by calculating the number of contributions for each commentor id.

frequency = df['commenter_id'].value_counts()
Chapter 4. Technical implementation details

Each feature is then added to the original training dataset.

    df['commentor_frequency'] = df['commenter_id'].apply(lambda x: 
        frequency[x]/df['node_id'].count())
    df['degree_centrality'] = pd.Series(degree_centrality)
    df['betweenness_centrality'] = pd.Series(betweenness_centrality)
    df['eccentricity'] = pd.Series(eccentricity)
    df['is_leaf'] = pd.Series(is_leaf)

4.3 Manual Labelling

As we saw in Chapter 3, a training set should be set aside from the main dataset for manual labelling. Given the heavy weighting of character count in our classifier's feature selection (as will be discussed in the following chapters), it may be possible to use the training dataset provided with this paper. However, if this dataset does not yield adequate results, then the following steps must be taken to select and label the training dataset.

We argue that the training set should be as large as necessary for the classifier to have an F1-score of above 0.8, while reducing the amount of manual labor for the researcher. As we will see, random forests performs well with small training sets, and as such, we obtained good results with a training set of approximately 700 comments to analyze a dataset of over 40,000 comments. The size of the total dataset, however, is not as much a factor in the size of the training set as the variety of data represented. The training data must be representative of the overall dataset, in order for the classifier to extrapolate features and weights in a relevant way.

Once selected, the training set should then be set aside and manually coded for the discourse quality index. The DQI code book used for this is described in Chapter 2, Table 4. Each comment should be manually coded using the guidelines in the code book, as such:
Chapter 4. Technical implementation details

Table 11: Manual DQI coding

<table>
<thead>
<tr>
<th>Comment</th>
<th>Participation</th>
<th>Level of Justification</th>
<th>Content of Justification</th>
<th>Respect</th>
<th>Counterarguments</th>
<th>Constructive Politics</th>
<th>Total DQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best score</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>Take back the voting restrictions</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Please cut down taxes for middle class</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Everything costs too much</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>&quot;We have dealt with EDUCATION, which of course is vital, and with MEASUREMENT and DATA, without which we'd be navigating without a</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>
From this, we calculate the total DQI for each comment by summing the columns. Our classifier will use DQI categories instead of exact scores, so we add a category score. The category for each comment should be 0=low quality (0-5), 1=medium quality (6-10), or 2=high quality (11-14).

Table 12: From DQI score to category

<table>
<thead>
<tr>
<th>Comment</th>
<th>Total DQI score</th>
<th>DQI category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take back the voting restrictions</td>
<td>5</td>
<td>Low</td>
</tr>
<tr>
<td>Please cut down taxes for middle class</td>
<td>7</td>
<td>Medium</td>
</tr>
<tr>
<td>Everything costs too much</td>
<td>5</td>
<td>Low</td>
</tr>
<tr>
<td>&quot;We have dealt with EDUCATION, which of course is vital, and with MEASUREMENT and DATA, without which we’d be navigating without a compass. However[...]&quot;</td>
<td>11</td>
<td>High</td>
</tr>
</tbody>
</table>

These scores should be appended in a column of the dataset, which for our purposes is called ‘DQI_groups’.

The columns are now as follows:

<table>
<thead>
<tr>
<th>comment</th>
<th>commentid</th>
<th>created_time</th>
<th>userid</th>
<th>username</th>
<th>parent_id</th>
<th>DQI_groups</th>
</tr>
</thead>
</table>

4 For the complete comment, see Appendix B
4.4 Classifier Training

Once the training dataset is obtained, the following steps are taken:

4.4.1 Split the labelled dataset

The training dataset should be split into two sets – a training set containing 80% of the data and a test set containing 20% of the data. This split should be done randomly.

```python
train, test = train_test_split(labelled_data, train_size = 0.8, random_state = 44)
```

4.4.2 Vectorize the words

Next, the comments should be vectorized, which will quantify the words into features. This is stored as a sparse matrix, as in the example below:

Comment 1: “I believe taxes should be lower.”
Comment 2: “Taxes are too high.”

<table>
<thead>
<tr>
<th>Feature</th>
<th>I</th>
<th>believe</th>
<th>taxes</th>
<th>should</th>
<th>be</th>
<th>lower</th>
<th>are</th>
<th>too</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Comment 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The training dataset is thus vectorized. The maximum features in our dataset is capped at 6,000, which speeds up processing time.
vectorizer = CountVectorizer(analyzer = "word", tokenizer = None, 
preprocessor = None, stop_words = None, 
    max_features = 6000)

Once this is done, additional quantitative features should be appended to the dataset. If the dataset contains data from multiple platforms (as we will show in the following chapter), it is recommended to add a feature for each platform, as in the example below. In our experience, discourse quality varies per platform, and the platform features carry considerable weight.

4.4.3 Create random forests object and tweak classifier parameters

Once the features are set up, the classifier should be trained using a random forests algorithm. First, the forest object is created. The parameters shown here are the ones that worked best for the dataset that we used. These parameters should be tweaked to improve accuracy.

    forest = RandomForestClassifier(n_jobs=4, n_estimators=30, 
criterion="entropy", warm_start=True, 
max_depth=17, bootstrap=False)

4.4.4 Fit training set

The forest object is then trained to predict the DQI score based on the features that were previously added.

    y,_ = pd.factorize(train['DQI_groups'])
    forest.fit(train_labelled_data, y)

4.4.5 Run trained classifier on test dataset

The trained classifier should now be tested on the test dataset, which is the 20% of the dataset that was previously set aside.
result = forest.predict(test_labelled_data)

4.4.6 Validate classifier performance

In order to assess accuracy and, if necessary tweak features or parameters, we recommend creating a confusion matrix and a classification report. The confusion matrix will show the number of comments that the algorithm placed in each category and will compare the predicted score with the actual score, which had been manually labelled. The classification report will provide precision, recall and f1 statistics of the classifier.

    output = pd.DataFrame(data={"actual_DQI": test["DQI_groups"],
"predicted_DQI": result})

    print('Confusion Matrix:
', pd.crosstab(output["actual_DQI"],
output["predicted_DQI"], rownames=['Actual'], colnames=['Predicted']))

    print( 'Classification Report:
', classification_report(output["actual_DQI"], output["predicted_DQI"]))

An example of the confusion matrix and the classification report are shown below:

Table 14: Example of classification report

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F-1</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.86</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>1</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
</tbody>
</table>
4.5 Classification

Once the classifier is trained, we can now analyse unlabelled data. This will be the remaining portion of the dataset that has not been manually given a DQI group score. Before beginning the analysis, it is important that it be prepared in the same way as the training dataset. We can now use both supervised and unsupervised methods to understand the data.

The trained classifier should be run on the unlabelled data. This will score each comment in the dataset as low (0), medium (1) or high quality (2).

```python
result = forest.predict(clean_unlabelled_data)
output = pd.DataFrame(data={"comments":
clean_unlabelled_data["comment"], "predicted_DQI": result})
```

**Table 15: Sample classifier predictions**

<table>
<thead>
<tr>
<th>Comment</th>
<th>Predicted DQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Prime Minister, I like your style.</td>
<td>0</td>
</tr>
<tr>
<td>Like @steve_smith, I believe that more investment in elder care would benefit everyone.</td>
<td>1</td>
</tr>
</tbody>
</table>

4.6 Clustering

A k-means clustering algorithm can now be run on the data. This will cluster the comments into topic groups.
Chapter 4. Technical implementation details

4.6.1 Vectorize 3-grams using TF-IDF

Analyze the data using term frequency-inverse document frequency (TF-IDF) which provides a ranking of the most important three-grams (phrases of three words) in the dataset.

```python
vectorizer = TfidfVectorizer(use_idf=True, analyzer='word',
ngram_range=(3,3))

tfidf_unlabelled_data = vectorizer.fit_transform(dataset[clean_unlabelled_data'])
```

4.6.2 Perform singular value decomposition (SVD)

Create a single-value decomposition object which will be used to factorize the feature matrix.

```python
svd = TruncatedSVD(n_components=100)
```

4.6.3 Normalize dataset

Create a normalizer object.

```python
normalizer = Normalizer(copy=False)
```

4.6.4 Perform latent semantic analysis (LSA)

Create a latent semantic analysis object that combines the two above transformations into a data workflow, then apply it to the dataset.

```python
lsa = make_pipeline(svd, normalizer)
lsa_transformed_data = lsa.fit_transform(tfidf_unlabelled_data)
```

4.6.5 Create k-means object

Create a K-Means object with 5 clusters (this can be changed if more topic groups are appropriate for the dataset), then apply the k-means algorithm to the transformed data.
An example of a key term clusters in a dataset would be as follows:

**Table 16: Sample clusters**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>moving edmonton community; processing center central; moved edmonton anything; question case processing; quite concerned community; reason moved edmonton; particular piece legislation; think important relationship; welcome thank coming; right community trying</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>mental health issues; struggles mental health; let talk day; question federal government; opening today bell; minister coming today; mental health initiatives; personally family sure; prime minister coming; many us room</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>mr prime minister; evening mr prime; want know miss; miss classroom think; minister education student; planning anything change; prime minister education; made canada strategy; question actually related; related carbon tax Cluster 3.; boil water advisories; many different ways; interests white people; young people young; talk young people; make sure truly; need youth center; people community know; minister took based; package time shifts</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>boil water advisories; many different ways; interests white people; young people young; talk young people; make sure truly; need youth center; people community know; minister took based; package time shifts</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>thank prime minister; overseas development assistance; predictably increase budget; oda overseas development; question predictably increase; question international development; recent years question; percent budget oda; role canada internationally; oda percent budget</td>
</tr>
</tbody>
</table>

### 4.7 Algorithm and parameter validation

Once the classifier and clustering algorithms are in place, they should be validated in order to ensure that they are set up optimally.
Chapter 4. Technical implementation details

4.7.1 Classification validation

The choice of classifier has been validated by comparing random forests with SVM and logistic regression and selecting the algorithm with the best results (random forests), as we will see in Chapter 5. The results themselves should be validated during the training phase, where 20% of the training dataset is set aside to validate the results of the classifier. As we saw above, the confusion matrix and classification report will allow a good sense of the accuracy of the results.

4.7.2 Clustering validation

In the validation of clusters, we have two options – external and internal validation. In our case, external validation is not possible because we have no training dataset in which the comments are already categorized (also called a ground truth). We therefore use internal validation. For k-means, we selected silhouette analysis.

The silhouette coefficient calculates the separation between clusters and the closeness of items within the cluster. A clustering output should have a silhouette coefficient between -1 and 1. If we continue on from our cluster analysis above, we add the following lines:

```python
silhouette_coefficient = metrics.silhouette_score(lsa_transformed_data, labels_, sample_size=1000)
```

If the silhouette coefficient is below 0, the comments may have been assigned to the wrong cluster. If the number is above 0 but still relatively low, this indicates that the clusters are close to each other and even overlapping. If the dataset is of relatively good quality, it may be possible to obtain a higher silhouette score by decreasing the cluster number, to reflect the topic variety in the discussion. As we will discuss in Chapter 7, we see closer clusters with higher discourse quality discussions, attributed to a convergence in discussion topics.

4.8 Interpretation

The DelibAnalysis framework provides two results: first, a prediction of discourse quality category for each comments; and second, a collection of 2 and 3-gram clusters that give context to the discussion.
4.8.1 What is the topic configuration of the conversation?

The following exercise shows each the most important terms for each cluster. This allows us to understand the topics under discussion.

```python
idf = vectorizer.idf_
idf_dict = dict(zip(vectorizer.get_feature_names(), idf))
idf_df = pd.DataFrame.from_dict(idf_dict, orient='index')
idf_df = idf_df.sort_values(by=0, ascending=False)
for i in range(true_k):
    df = pd.DataFrame(columns = ['ngram', 'TF-IDF-score'])
    for ind in order_centroids[i, :15]:
        df2 = pd.DataFrame([[terms[ind],idf_df.get_value(index=terms[ind], col=0)]],
                           columns = ['ngram', 'TF-IDF-score'])
        df = df.append(df2)
    df = df.sort_values(by='TF-IDF-score', ascending=False)
    plt = df.plot(kind='barh', legend=None, color='#33cc33', edgecolor='black', linewidth=1.0,)
    plt.invert_yaxis()
    plt.set_yticklabels(df['ngram'])
    plt.set_title("Top 15 3-grams by importance: Cluster " + str(i))
    plt.set_xlabel("TFIDF Score (importance in dataset)")
    plt.set_ylabel("3-grams")
```

This produces a series of bar charts such as the one below:

![Bar chart showing top 15 3-grams by importance for Cluster 5](image)

We then examine the distribution of the comments over the clusters, in order to understand if there were topic clusters that were more important than others.
lab = pd.DataFrame(labels)
plt = lab.plot(kind='hist', legend=None, bins=np.arange(16)-0.5, 
edgcolor='black', linewidth=1.0, xticks=range(0,15))
plt.set_title("Comment count per cluster")
plt.set_xlabel("Cluster Number")
plt.set_ylabel("Count")

In the example below, Cluster 10 only has 9 comments, whereas Cluster 0 has 91 comments. Therefore, in this conversation, topic 0 is the most extensively discussed.

4.8.2 What is the DQI category distribution of the conversation?

This question allows us to understand the DQI quality of the conversation as a whole. We use a histogram and output basic statistics to calculate the probability that a given comment has a low, medium or high DQI category.

plt = output['predicted_DQI'].plot(kind='hist', legend=None, 
bins=np.arange(4)-0.5, edgcolor='black', linewidth=1.0, 
xticks=range(0,3))
plt.set_xlabel("Predicted DQI")
plt.set_ylabel("Comment Count")
plt.set_xticklabels(['low', 'medium', 'high'])
plt.set_title("Trudeau Saskatchewan Comment Count by Predicted DQI")
print('Mean', output['predicted_DQI'].mean())
print('Minimum', output['predicted_DQI'].min())
Chapter 4. Technical implementation details

print('Maximum', output['predicted_DQI'].max())
print('Standard Deviation', output['predicted_DQI'].std())
print('Skew', output['predicted_DQI'].skew())
print('Kurtosis', output['predicted_DQI'].kurt())
print('\nProbability of a comment being scored each category')
print(output['predicted_DQI'].value_counts(1))

This gives us the following chart and descriptive statistics:

![Chart with descriptive statistics]

4.8.3 What are the important parameters for calculating the DQI category?

Finally, we examine the parameters, or features, that were most important in training our classifier to accurately predict the discourse quality of our data. In the code below, we obtain the importances and plot the top 10 features by order of importance.

```python
importances = forest.feature_importances_
std = np.std([tree.feature_importances_ for tree in forest.estimators_],
             axis=0)
```
This is of particular interest to us as we examine what features in a platform might be more conducive to democratic quality.

4.9 Summary

In this chapter, we provided technical implementation details for each step of the DelibAnalysis framework detailed in Chapter 3. We gave code snippets in Python for each step, along with relevant data samples and charts, as required. The complete implementation of the code is provided in the Appendix. The following chapter will detail the experimental process that led to the definition of the DelibAnalysis framework.
5. Design considerations

This chapter explains the steps that were taken to develop the Framework provided in Chapters 3 and 4. It details the design choices that were made, particularly when it came to classification model selection and the combination of structured and unstructured analysis.

5.1 Classifier selection

A classifier was first trained using the 720 labelled comments in the dataset. These comments were manually annotated using the DQI coding manual (Steenbergen et al, 2003). As we have seen, the DQI model by Steinbergen et al (2003) used six indicators (participation, level of justification, content of justification, respect, counterarguments and constructive politics) in order to construct a discourse quality index.

When using machine learning, however, we hypothesized that the DQI could be arrived at by completely different indicators, or features, as they are referred to in machine learning. When the DQI is calculated using the six indicators, it can have a minimum score of 0, or a maximum of 14 (the sum of the highest scores for each indicator). Rather than using a numeric value, we separated the numeric score into three categories (low quality, medium quality and high quality).

We then separated the dataset into a training set consisting of 80% of the comments and a test set consisting of 20% of the comments. The training set would be used to build the predictive model and the test set would be used to evaluate its accuracy.

The first category of features that was used as part of the predictive model was bag of words. It consists of converting the words in each comment into a separate feature, with a score of 1 if the word was in the comment and a score of 0 if it was absent. This generated a large sparse matrix of word features for the entire training set - 6010 word features for 576 comments. We then used this data to train three supervised machine learning classifiers – Random Forests, Support Vector Machines (SVM) and Logistic Regression. These will each be described in turn below.
5.1.1 Random Forests classification

Random forests is a type of machine learning algorithms called ensemble learning, in which a series of separate classifiers predict based on a random subset of features. The results from these classifiers are then aggregated into a more robust prediction.

Each individual classifier is a decision tree algorithm, which uses a tree structure to arrive at the final prediction based on the features that have been provided. Because decision trees have a tendency to overfit the data and go too deep, particularly when a lot of features are involved, the random forests method allows for a faster and more consistent implementation (Python SciKit Learn, 2017). The random forest classifier’s parameters are listed in Appendix A.

In the example below, decision-tree diagram shows the decision path for one tree and a random set of features in order to decide between low and high DQI. In our implementation, the random forest would contain 100 such decision-trees, aggregating the results to obtain the classification decision.

Figure 7: Example decision-tree
Chapter 5. Design considerations

5.1.2 Support Vector Machine (SVM) Classification

An SVM is a classifier algorithm which separates data into classifications based on an optimal separating hyperplane. This separation between different classes is surrounded by a margin, which according to the algorithm should be as wide as possible given the data. Given a training dataset, the SVM creates an optimal separation so that data with are associated with their classification on either side of the line. Non-labelled data is then assigned to each side based on their weight/feature combinations. When there are multiple classifications to assign (as in the case of assigning eye colour to green, blue, brown or black), the SVM can in fact be used in an infinite-dimensional space (Python SciKit Learn, 2017).

5.1.3 Logistic Regression Classification

Despite its name, logistic regression is a linear classification rather than a regression model. In this mode, the probabilities of data being assigned to classifications are modelled using a logistic function. The basic logistic regression model assigns the data to two classifications (0 or 1). Our model uses multinomial logistic regression - that is, using probability to determine one of multiple categories (Python SciKit Learn, 2017).

5.1.4 Results of the classifier selection process

As we will see in the table below, the random forests algorithm performed significantly better than SVM and logistic regression. This is coherent with related research, described further in section 5.5. As we saw in section 3.3.7, we used a confusion table for each algorithm to measure precision, recall and F1-score, and selected the method that yielded the best F1-score.

Table 17: Classification report for random forests classifier

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>84</td>
</tr>
<tr>
<td>Medium</td>
<td>0.72</td>
<td>0.78</td>
<td>0.75</td>
<td>23</td>
</tr>
</tbody>
</table>
Chapter 5. Design considerations

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.84</td>
<td>0.87</td>
<td>0.85</td>
<td>84</td>
</tr>
<tr>
<td>Medium</td>
<td>0.42</td>
<td>0.43</td>
<td>0.43</td>
<td>23</td>
</tr>
<tr>
<td>High</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>5</td>
</tr>
<tr>
<td>Average/Total</td>
<td>0.71</td>
<td>0.74</td>
<td>0.73</td>
<td>112</td>
</tr>
</tbody>
</table>

Table 18: Classification report for SVM classifier

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.85</td>
<td>0.88</td>
<td>0.87</td>
<td>84</td>
</tr>
<tr>
<td>Medium</td>
<td>0.50</td>
<td>0.52</td>
<td>0.51</td>
<td>23</td>
</tr>
<tr>
<td>High</td>
<td>1.00</td>
<td>0.20</td>
<td>0.33</td>
<td>5</td>
</tr>
<tr>
<td>Average/Total</td>
<td>0.79</td>
<td>0.78</td>
<td>0.77</td>
<td>112</td>
</tr>
</tbody>
</table>

Table 19: Classification report for logistic regression classifier

5.2 Selection of candidate classification algorithms

The three candidate classification algorithms, SVM, logistic regression and random forests, were all selected because of their ease of interpretation. As opposed to deep neural network classifiers, for example, these more "traditional" classifiers output a much clearer relationships between weights and features, which is critical for our analysis. Of the three supervised learning machine algorithms, SVM appear to be the most commonly used in content analysis research, followed by logistic
regression and by random forests. SVM have often been used for sentiment analysis, with two (positive or negative) or three (positive, neutral or negative) categories. Users of the algorithm have tended to argue that it is well suited for sentiment analysis or for topic analysis (Mullen and Collier, 2004; Pang et al, 2002). Diedrich et al (2003) used SVMs for authorship detection of texts, arguing that it is more appropriate than other methods of classification for its ability to process hundreds of thousands of features (due to its infinite dimensionality). Mitra et al (2007) compared SVM to two other classification algorithms - K Nearest Neighbors and Naive Bayes - in order to classify noisy document titles, and find the SVM outperformed the other two in all tests.

In turn, logistic regression, while less common for content analysis, nevertheless is used with interesting results. Several authors have used logistic regression to predict use of semantic or syntactic elements. Fiss and Hirsch (2005), for example, used logistic regression to predict the occurrence of discourse on globalization in three major newspapers. Arnold et al (2000), on their part, used the classifier to predict the use of various sentence structures in order to understand the circumstances under which one might be used over others. Allen and Lincoln (2004) examined the relationship between critical discourse about a film and their subsequent recognition as great films.

Finally, random forests classification has been used in discourse quality analysis. Blanchard (2015) and Wang (2014), for example, used the algorithm when examining teacher and student speech in the classroom. Stab (2014) used it to identify discourse structures in argumentative essays. Random forests have also been used specifically for social media analysis on a number of occasions (Can 2013; De Choudhury 2012; Lee 2011; Wagner 2012). The popularity of the algorithm is due largely to the fact that it is easy to implement, quick to train and versatile (El Deeb, 2015). In their empirical evaluation of supervised learning algorithms, Caruna et al (2008) rank random forests first in terms of performance, ahead of neural nets, boosted trees and SVMs. Random forests have a number of well-known features, including accuracy and large number of variable handling (useful because we use large sparse matrices for word analysis) (Berkeley, 2017).

Blumenstock (2009) also found that bag of words analysis produced a maximum of 81% accuracy when measuring the quality of Wikipedia articles. However, additional refining of the classifier may allow for an improvement in the accuracy of the tool. When comparing random forests, SVM and logistic regression, as we have seen, random forests had better accuracy – 88% instead of 71% for SVM and 79% for logistic regression.
Therefore, we selected the random forests classifier in order to classify unlabelled data.

5.3 Selection of k-means clustering algorithm

In the classification exercises, it is clear that the majority of comments used were of low discourse quality. However, upon manual inspection, it appeared that they were thematically relevant. That is, although the quality of discourse might have been low, the topics mentioned in nearly every post were politically relevant. Topics that arose, for example, included immigration, mental health, agriculture, hydroelectricity, minimum wage, taxes, defense, oil, carbon tax, the working poor, the UN, globalization, the environment, retired people and pipelines.

While random forests, SVM and logistic regression are supervised classification algorithms, k-means is an unsupervised clustering algorithm. This means that it requires no training data, and instead of classifying data into three pre-determined categories, it will create numbered clusters that have similar features. K-means is often used to analyze social media content. Priem (2012) capture scholarly impact “flavours” in five groups across social media, for example. Cranshaw (2012) remap city neighbourhoods according to social media k-means clusters. In discourse analysis more broadly, topic clusters are a commonly used technique to understand language patterns, whether it be education (Hovardath 2006), newspapers (Pollack 2011), or student’s online discourse quality (Chiu, 2010).

TF-IDF is commonly used in natural language processing. Robertson (2004), for example, discusses the theoretical justification for using the measure. It is used primarily to determine the most important terms in a document, which may be different from the most common. If a term is frequent in a document but infrequent in the whole dataset, it will weigh much more heavily as a feature than if it is both frequent in a document and in the whole dataset. This also helps to identify stop words, such as the, and, or, which might be frequent but have no analytical value.

5.4 Summary

In this chapter, we discussed the design considerations that led us to the development of the DelibAnalysis framework. We detailed the classification selection process, which gave us the best
performing algorithm from SVM, logistic regression and random forests. We described the rationale for the use of the candidate classification algorithms as well as the use of k-means in unsupervised analysis. The following section will apply the DelibAnalysis framework on a mixed dataset extracted from political blogs and Facebook deliberations.
6. Instantiation of framework

This chapter instantiates the analytical steps outlined in Chapters 3 and 4 using real world examples. The objective of this is to demonstrate the validity of the framework by showing its application in a variety of settings. In order to achieve this, we selected a mixed dataset, consisting primarily of blog and Facebook data, as well as a live Townhall recording for comparative purposes. We also instantiated the framework using node-level features, using a United Nations deliberation on poverty reduction policy. In the following sections, we will go through the steps of the DelibAnalysis framework and touch on the interpretation of results, which will be further developed in Chapter 7. In order not to duplicate information, we refer to these two instantiations as Mixed Instantiation and Node Instantiation. Steps that are the same for both, such as data collection and processing, will be discussed together.

6.1 Data collection and processing

The data selected represented a range of political conversations on Facebook Live, on political blogs and during live Townhalls. A total of 78,726 comments were collected from various discussions, 720 of which were manually coded using the DQI coding manual in order to provide labelled data to train the classifiers. The source of each data is listed below, along with examples from each dataset.

<table>
<thead>
<tr>
<th>Name of dataset</th>
<th>Desc.</th>
<th>Start date</th>
<th>Total comments</th>
<th>Comment s used for training</th>
<th>Method of collection</th>
<th>Instantiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernie Sanders</td>
<td>Facebook live comments from the streaming of March 1, 2017</td>
<td>25,581</td>
<td>100</td>
<td>Facebook API</td>
<td>Mixed</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 6. Instantiation of framework

| an interview with Senator Bernie Sanders (VT) |
| "This is the Man who should be leading the Free World!"
| "I'm 45. This was the first time I ever registered to vote because I wanted to see someone with integrity in the White House. I hope you run again in four years!"
| "Future president of the Socialist State of America???

| Republican Townhalls | Facebook live comments from the streaming of a series of in-person town halls with Republican Senator Pat Toomey (PA), Representative Scott Perry (PA) and Representative Mark Sanford (SC) | Febrary 16, 2017; Febrary 18, 2017; March 6, 2017 | 6,952 | 100 | Facebook API | Mixed |

| "The Pell Grant system has not worked very well. It’s driven up the cost of college education and left students in debt. There are better ways”  
| "I’m actually impressed by how much this discussion is less talking point bs and, while it is tense, it is substantive and honest, and I would say fairly civil.” |
Chapter 6. Instantiation of framework

<table>
<thead>
<tr>
<th><strong>Justin Trudeau Online Townhalls</strong></th>
<th><strong>Facebook live comments from the streaming of a series of in-person town halls in Kingston (ON), London (ON), Saskatoon (SK), Sherbooke (QC) and Yellowknife (NWT)</strong></th>
<th><strong>January 13, 2017; January 15, 2017; January 16, 2017; January 17, 2017; January 18, 2017; January 26, 2017</strong></th>
<th><strong>31,764</strong></th>
<th><strong>100</strong></th>
<th><strong>Facebook API</strong></th>
<th><strong>Mixed</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EULive</strong></td>
<td><strong>Facebook Live event in advance of June 10, 2016</strong></td>
<td><strong>June 10, 2016</strong></td>
<td><strong>14,099</strong></td>
<td><strong>100</strong></td>
<td><strong>Facebook API</strong></td>
<td><strong>Mixed</strong></td>
</tr>
</tbody>
</table>

“How do you think the US and the current administration should deal with the REAL threats to our security-- meaning the influence of Russia and Putin on our government officials?”

“What about the pipeline you gave the ok. That is a nightmare waiting to happen. Without clean water clean air and the fracking of our planet we no longer exist. We need you to use common sense on this issue. PLEASE don’t let anyone or anything destroy the human race.”

“Allons-y tout! Devenons tous bilingues, un enrichissement pour tous. Let’s go! Let’s become all bilingual, a richness for all!”

“More renewable energies, less fossil energy”
| Justin Trudeau in-person Townhall | Transcript from an in-person Townhall with Justin Trudeau in Saskatoon, Saskatchewan. | January 26, 2017 | 26 | 26 | Manual transcription of video recording | Mixed |

“Thank you thank you Saskatoon. Thank you Ralph for your warm words, thank you President for welcoming us here, thank you for the welcome on traditional Treaty 6 territory, it’s a real pleasure to be here. Please, sit down, we’re all friends here.[…]”

“Good evening Mr. Prime Minister, thanks for coming. The question that I have for you actually it’s related to the carbon tax policy.[…]”


“Does anyone have an idea of what would happen if the amendments to the constitution are not passed? […]”
“I also agree with postponing the Biennial Convention and leadership vote until at least the dates specified in the amendment. [...]”

“Hello, I fully support the proposed amendments. I believe the time frames laid out make sense. The Party needs time to regroup and refocus before choosing the right person to lead us in the next elections.”

<table>
<thead>
<tr>
<th>UNDP Poverty Reduction consultation comments</th>
<th>Comments for a moderated open consultation on the Sustainable Development Goals, and specifically poverty reduction with UNDP.</th>
<th>March 16, 2017</th>
<th>170</th>
<th>170</th>
<th>Manual scraping from website</th>
<th>Node</th>
</tr>
</thead>
</table>

“To add to the notes on adjusting policies to support co-operatives, I would like to highlight the specific ways in which co-operatives serve as tools for eradicating poverty and how the UN and its Member States could support their development in a practical way. [...]”

“In order to achieve such goal, we will need the cooperation of governments and commitment. That is since most causes of problem is related to governmental policies and high level corruption.[...]”

“We do have plethora of literature on poverty measurement as well as measures for eradication. However, certain points need to be outlined.[...]”

| Total | NA | NA | 78,726 | 720 |
In the case of the Facebook comments, the data was collected using the Facebook Developer API during a period from December 2016 to February 2017. The in-person Townhall data was collected by downloading the full-length video which was shared on Facebook and transcribing the entirety of the discussion manually. The UNDP poverty reduction comments were scraped manually from the platform. The Liberal Party Renewal Blog comments were pre-labelled in Fournier-Tombs (2013), with an inter-rater agreement score of 0.72. The data was then processed as described in section 4.2.2.

6.2 Adding features

As part of processing, we added two quantitative features – comment category and character count. The distribution in the dataset for both features are described below. Only the UNDP Poverty Reduction consultation comments were appropriate for graph structure and node-level features; we will describe this analysis further in section 6.4.

6.2.1 Distribution of comment category

The comments were extracted from three different type of platforms – Facebook, blogs (as comments to an original post), and live Townhall (transcribed from the publicly available recording). The comment count per category is charted below.

![Comment Category by Count](image)

*Figure 8: Comment count by category in the training dataset*

6.2.2 Distribution of character count

The distribution of the length of the comments used in the training sample is charted below. Comments tended to be under 1000 characters, with some exceptions.
6.3 Manual labelling

The training dataset was manually annotated on the following features: discourse quality (participation, level of justification, content of justification, respect, counterarguments and constructive politics). Although, as mentioned above, the classifier aggregated these scores and look only for three overall discourse quality scores - low, medium and high, the data was labelled using the original DQI coding manual, giving a total of 0-14 on the DQI scale. As we have discussed, we followed the DQO code book first introduced in Chapter 2, Table 4. These scores were later normalized during the building of the classifier.

The distributions of the scores for each feature in the training dataset are listed as follows.

6.3.1 DQI Indicator 1: Participation

There were two possible participation scores - 0 (no participation) and 1 (participation). Comments were given a 0 when they consisted in an interruption of another commenter. In this case, as is the case with most online discussions, the barrier to participation was low. As we will see below, interruptions were rare, and the majority of comments will receive a participation score of 1. However, in cases where participants actively tried to censor each other by saying disrespectful words or asking each other to leave, we gave a 0.
6.3.2 DQI Indicator 2: Level of justification

There were four possible level justification scores: no justification (0), inferior justification (0.33), qualified justification (0.67) and sophisticated justification (1). Justifications were examples used to add weight to an argument. The more justifications were included, the higher the score. As we can see below, more comments had no justification at all, however, there were a number of examples in each category.
6.3.3 DQI Indicator 3: Content of justification

This indicator examined the quality of the justification. The possible scores were: explicit statement concerning group interests (0), neutral statement (0.33), explicit statement of the common good in utilitarian terms (0.67), explicit statement of the common good in terms of the difference principle (1). There were many comments that related to group interests, for example, commenters bringing forward their own group’s requests – better pension for the retired, lower tuition for the youth, reparation for aboriginal populations, and so on. Then, with a score of 0.33, many commenters did not make any reference to a justification at all; this was scored as a neutral statement.

![Comment count by content of justification score in the training dataset](image)

*Figure 12: Comment count by content of justification score in the training dataset*

6.3.4 DQI Indicator 4: Respect

There were three possible scores for respect - no respect (0), implicit respect (0.5) and explicit respect (1). Typically, no respect included expletive or disparaging words directed at another speaker. Explicit respect, on the other hand, will include praise or words of thanks to another speaker. Comments not containing either category were categorized as implicit respect. As you will see, the majority of comments were categorized as implicit respect.
Chapter 6. Instantiation of framework

6.3.5 DQI Indicator 5: Counterarguments

There were four possible scores for the counterarguments indicator. Counterarguments ignored (0) - when the commenter did not include any other arguments in the comment, counterarguments included but degraded (0.33), counterarguments included - neutral (0.67), and counterarguments included and valued (1). The majority of comments in our dataset did not include any counterarguments.

6.3.6 DQI Indicator 6: Constructive Politics

Three scores were possible in constructive politics. Positional politics (0) - when a or commentor stuck to their position with no attempt at compromise, alternative proposal (0.5) - when the
commenter made a comment that provided another perspective without including other comments in the proposal, and mediating proposal (1), where the commenter included several perspectives in an attempt to form a conciliating position. Most of the comments stated their own position without an attempt to compromise.

Figure 15: Comment count by constructive politics score in the training dataset

Once the manual coding was completed, we combined the scores to find the DQI score, and then converted the score to a DQI category of low, medium or high.

6.3.7 Classifier training

As we discussed, we split the labelled dataset into a training (80%) and testing (20%) set. We therefore had 556 comments in the training dataset and 112 comments in the testing dataset. We then vectorized the words in the dataset to create a sparse matrix structure. We created a random forests object with the default DelibAnalysis parameters described in Chapter 4, and fit our training dataset to the forest. The classifier was then run against the testing set, with the following results.
Table 21: Classification report

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>86</td>
</tr>
<tr>
<td>Medium</td>
<td>0.79</td>
<td>0.82</td>
<td>0.80</td>
<td>55</td>
</tr>
<tr>
<td>High</td>
<td>1.00</td>
<td>0.33</td>
<td>0.50</td>
<td>3</td>
</tr>
<tr>
<td>Average/Total</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>144</td>
</tr>
</tbody>
</table>

Table 22: Predicted vs actual report

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>76</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium</td>
<td>10</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Actual high</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

In our implementation, we also calculated the Pearson chi-square\(^5\) value to observe the statistical significance of the classifier. We interpreted the null hypothesis as equivalent to guessing; a 3-category random classifier would be correct 33.33% of the time, or for 48 observations. When comparing observed to expected (null) values, obtained a chi-square value of 142.18, well above the critical chi-square value for a p-value (statistical significance) of 0.5%, which is 3.84.

For additional validity, we calculated the multivariate McNemar-Bowker value, which compared the accuracy of the classifier on the testing dataset with the accuracy of the manual training,

\[\chi^2 = \sum_{i=1}^{k} \frac{(o-e)^2}{e}\]
which we evaluated as accurate 100% of the time. This yielded a chi-square value of 2, which gave a p-value of 0.57%, on the border of statistical significance.

With these tests, concluded that the classifier provided statistically significant additional value and decided to move ahead with the analysis.

Although there were 4,285 features in this model, the top 50 features ranked by importance are shown as follows.

Figure 16: Top 50 features by importance for random forests classifier using hybrid features
6.4 Classification and clustering

The random forests classifier and the k-means TF-IDF algorithm were then used to evaluate non-labelled datasets from Facebook live conversations. In this instantiation, we decided to separately classify each of the sub-sets of data described in table 10, in order to better compare between datasets. We also did the clustering exercise for each subset in order to understand the context specific topics in each case. The results from these classification and clustering exercises are below.

6.4.1 Bernie Sanders dataset analysis

For this dataset, the clustering silhouette coefficient was 0.122 – indicating relatively close clusters. The vast majority of comments received a low DQI score. The clusters were not so even, with a high amount of comments concentrated in cluster 6. The topic configuration was interesting, with relevant topics such as social security and medicare surfacing.

![Figure 17: Discourse quality prediction (random forests) for Bernie Sanders dataset](image)

![Table 23: K-means clusters based on TF-IDF for Bernie Sanders dataset](image)
6.4.2 Republican Townhalls dataset analysis

Again, in this dataset, the majority of comments received a low DQI score. The clusters were unevenly distributed, with many of the comments concentrated in cluster 3. The silhouette coefficient was 0.130, which indicates more diversity than in the Bernie Sanders dataset. The topic clustering showed a lot of very current themes, including the US Environmental Protection Agency (EPA), the Keystone Pipeline and the Russian interference with the US elections.

Figure 18: Discourse quality prediction (random forests) for Republican Townhall dataset
6.4.3 Justin Trudeau Townhalls dataset analysis

In this dataset, we see again low DQI scores. The clusters were unevenly distributed with a spike in cluster 3. The silhouette coefficient, tellingly, was 0.141. Although slightly higher than the previous datasets, the topic clusters were close together, with a lot surfacing on carbon taxes and not much on other topics.
Table 25: K-means clustering based on TF-IDF for Justin Trudeau Townhalls dataset

6.4.4 EULive dataset analysis

The EULive dataset was of the poorest quality, with a silhouette coefficient is 0.308. Again, the majority of comments were of low quality, and the clusters were unevenly but more distinctly distributed, showing only in favour or against positional statements, with little attempt at convergence.
Chapter 6. Instantiation of framework

Figure 20: Discourse quality prediction (random forests) for EULive dataset

Table 26: K-means clusters based on TF-IDF for EULive dataset

The Facebook comment predictions are largely skewed towards a score of 0, or “low” discourse quality. This is expected, especially if one takes into consideration low character count. The clustering analysis, however, tells a slightly different story, especially when one looks at Cluster 1 in the Trudeau Town Hall dataset (nodapl nokxxl unplugwynne keepyourpromises sweeney shelley alhamdulillah pme blury nour). “Nodapl” and “nokxxl” are presumably hashtags, used in reference
to the DAPL and Keystone XL pipelines, that have been largely discussed in the public sphere in terms of environmental consequences and land rights.

As a mode of comparison, we analyzed the 26 comments from one of Justin Trudeau's live Townhalls in Saskatoon, Saskatchewan. These comments were made verbally and were transcribed.

6.4.5 Trudeau Saskatchewan Townhall analysis

As can be seen in the counterexample above, the Trudeau Saskatchewan Townhall fares better in terms of discourse quality than the Facebook discussions. While there are some low and medium comments, the majority of comments were of high quality. The clusters were also very close together, with a silhouette coefficient is 0.002. Interesting and cohesive topic clusters included a topic on Overseas Development Assistance (ODA), and one on the effect of the closing of a case processing center in rural Alberta on local employment.

![Probability of a comment being scored each category](image)

**Figure 21: Discourse quality prediction (random forests) for Trudeau Saskatchewan townhall dataset**
Chapter 6. Instantiation of framework

Table 27: K-means clustering based on TF-IDF for Trudeau Saskatchewan dataset

6.5 Node instantiation

In the second instantiation, we examined the discussion network of the UNDP Poverty Reduction dataset. The data used for this analysis was sourced from the United Nations Development Programme (UNDP)’s Global Development Hub, a platform used to allow for deliberation on development and poverty reduction themes with the objective of including the results of the discussion in the organization’s various policies and position papers. According to the website, the “purpose of the 2017 ECOSOC electronic discussion (e-Discussion) is to engage stakeholder groups, experts, practitioners and policy-makers from various regions in a global dialogue on specific aspects of the 2017 ECOSOC theme.”

The discussion used, entitled “Eradicating poverty and leaving no one behind”, was initiated by a moderator and contains 197 comments. The structure of the discussion is an ego network, where the initial post is the center, or root of the graph, and the majority of comments are leafs, connected
only to the graph. However, it is possible for commenters to respond to other posts than the root post, and there are therefore also some secondary centers branching off from the root.

Two exercises were conducted. First, statistics and visualizations for the network as a whole were computed. The objective of this was to develop reliable indicators about the discourse quality of a conversation *as a whole*, based on its network statistics.

The three classification algorithms were then trained again, using the same 80/20 training/testing split as in Step 1. However, instead of using bag-of-words and numeric features, we used the node-level features computed in this phase.

Network-level measures were also captured, although they were not included in the analysis (which uses node-level measures). This was done in order to frame the results of the analysis. As most blog-based deliberations begin with an original post, they are all structured as ego networks, with varying centrality and depth measures.

Table 28: Network-level measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value for analyzed Data</th>
<th>Description</th>
</tr>
</thead>
</table>

![Network Structure with Leaf Colors](image1)

![Network Structure with Level Colors](image2)
Chapter 6. Instantiation of framework

<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Ego Network</td>
<td>True</td>
<td>Boolean</td>
</tr>
<tr>
<td>Average Degree Centrality</td>
<td>0.0118343195266</td>
<td>Double</td>
</tr>
<tr>
<td>Average Betweenness Centrality</td>
<td>0.00847877612583</td>
<td>Double</td>
</tr>
<tr>
<td>Average Cardinality</td>
<td>1.99411764706</td>
<td>Double</td>
</tr>
<tr>
<td>Diameter</td>
<td>4</td>
<td>Int</td>
</tr>
<tr>
<td>Average Clustering</td>
<td>0.0</td>
<td>Double</td>
</tr>
<tr>
<td>Degree Pearson Correlation Coefficient</td>
<td>-0.636600591151</td>
<td>Double</td>
</tr>
<tr>
<td>Number of triangles</td>
<td>0</td>
<td>Int</td>
</tr>
</tbody>
</table>

Secondly, statistics at the node level were computed. The objective of this was to develop reliable indicators about the discourse quality of *a given node*, based on its position in the network.

The following node-level measures were selected:

*Table 29: Node-level measures*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Sample value for analyzed node</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree centrality</td>
<td>0.786982</td>
<td>Double</td>
<td>A measure of how many</td>
</tr>
</tbody>
</table>
The above visualizations and statistics were computed using the NetworkX Python library, which allows for the creation of graph objects, their visualization and associated computations (NetworkX, 2017).

The node-level features were used to train a new classifier, using Random Forests. The results are shown below.

### 6.5.1 Random Forests classification using node-level features

*Table 30: Classification report using random forests classifier and node-level features*

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Medium</td>
<td>0.89</td>
<td>0.80</td>
<td>0.84</td>
<td>10</td>
</tr>
<tr>
<td>High</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Average/Total</td>
<td>0.74</td>
<td>0.67</td>
<td>0.70</td>
<td>12</td>
</tr>
</tbody>
</table>
Chapter 6. Instantiation of framework

Table 31: Prediction vs actual report using random forests classifier and node-level features

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium-low</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Actual high</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In this dataset, we didn’t find that the node features had a strong relationship with discourse quality. However, when we conducted the clustering portion of the analysis, we obtained very promising results, with a good distribution among the clusters and very relevant topics. The silhouette coefficient was 0.024, which indicates similar clusters.

Figure 22: Top node-level features by importance
When we ran the previously Mixed classifier on this dataset, which as we have seen was much more effective, we obtained the following DQI category distribution.

6.6 Summary

In this section, we described the implementation of the DelibAnalysis framework on a mixed dataset and on a graph-represented dataset. We found that the graph-represented dataset, although yielding tepid results in node-only classification, had very strong clustered topics, much
better than any of the discussions obtained from Facebook. In the following chapter, we will further discuss our analytical results and situate them in the literature.
7. Further discussion of results

The following chapter further discusses the results of the application of the methodological framework on mixed and node datasets detailed in the previous chapter. This provides not only insight into deliberation online, but also serves to demonstrates how the framework in chapter 3 can yield interesting results when applied to real world data. After examining the results obtained from running the first classifier as well as the unsupervised clustering algorithm on our unlabelled comment, we discuss the most important features selected by the classifier. In so doing, we follow the three steps of the interpretation of results section of the DelibAnalysis framework. We conclude the chapter by proposing design features, based on our findings, that might make for conversations of higher democratic quality.

7.1 DQI category distribution

First, we examine the results of the discourse quality classifier on our unlabeled data. As we saw in Chapter 6, all Facebook discussions received low discourse quality scores. The UNDP Poverty Discussion, which was on a blog platform, as well as the counterexample, which was the Trudeau Saskatchewan Townhall, received much higher discourse quality scores. It is perhaps intuitively unsurprising that the Facebook comments would receive lower scores than engaged deliberative processes such as an online consultation and a live town hall. However, does this mean that the discussions were of no democratic value?

Freelon (2010), based on the work of Habermas and Dalhberg, argued that online discourse is of diverse nature, and as such should be evaluated in different ways. He outlined three categories of online content: liberal individualist, communitarian and deliberative. The characteristics of each category are listed below.
Chapter 7. Discussion of results

Table 32: Online content categories (Freelon, 2010)

<table>
<thead>
<tr>
<th>Online content category</th>
<th>Characteristics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal individualist</td>
<td>Monologue; personal revelation; personal showcase; flaming</td>
<td>Personal blog or article; comments using personal terms as an example (for example using illness as an argument for healthcare); promotion of a personal website; strong-worded criticism (as in a news website comment)</td>
</tr>
<tr>
<td>Communitarian</td>
<td>Ideological fragmentation; mobilization; community language; intra-ideological questioning; intra-ideological reciprocity</td>
<td>Social media platform (especially private group ones such as Facebook where each person has their own select network)</td>
</tr>
<tr>
<td>Deliberative</td>
<td>Rational-critical argument; public issues focus; equality; discussion topic focus; inter-ideological questioning; inter-ideological reciprocity</td>
<td>Targeted discussion or debate, such as online Townhalls (can be Facebook, a blog, or another platform)</td>
</tr>
</tbody>
</table>

Due to the asynchronous nature of online discussions, it can be argued that even a monologue can contribute to democratic deliberation by providing information and viewpoints that can then be debated on other platforms. When one looks, for example, at the Bernie Sanders discussion dataset analyzed in chapter 4, one sees that in the over 25,000 comments collected, there could be examples in each category. However, according to the way in which the DQI is constructed, the liberal individualist and the communitarian comments would receive much lower scores than the deliberative comments. The reason for this is broken down in the following table.

Table 33: Application of online content categories to Discourse Quality

<table>
<thead>
<tr>
<th>Online content category</th>
<th>Discourse Quality Indicator</th>
<th>Description</th>
</tr>
</thead>
</table>

92
### Chapter 7. Discussion of results

<table>
<thead>
<tr>
<th></th>
<th>Participation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal individualist</td>
<td>In the case of “flaming” or &quot;strongly worded criticism&quot;, might receive a 0.</td>
<td></td>
</tr>
<tr>
<td>Level of justification</td>
<td>May receive a 2 (qualified justification), or a 3 (sophisticated justification), depending on the level of detail.</td>
<td></td>
</tr>
<tr>
<td>Content of justification</td>
<td>Likely a 0 (explicit statement concerning group interests).</td>
<td></td>
</tr>
<tr>
<td>Respect</td>
<td>Either score (no respect, implicit respect, explicit respect).</td>
<td></td>
</tr>
<tr>
<td>Counterarguments</td>
<td>Most likely a 0 (counterarguments ignored).</td>
<td></td>
</tr>
<tr>
<td>Constructive Politics</td>
<td>A 0 (positional politics) by definition.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Participation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Communitarian</td>
<td>Most likely a 1</td>
<td></td>
</tr>
<tr>
<td>Level of justification</td>
<td>Most likely a 0 (no justification) or a 1 (inferior justification).</td>
<td></td>
</tr>
<tr>
<td>Content of justification</td>
<td>Most likely a 0 (explicit statement concerning group interests).</td>
<td></td>
</tr>
<tr>
<td>Respect</td>
<td>Most likely a 1 (implicit respect).</td>
<td></td>
</tr>
<tr>
<td>Counterarguments</td>
<td>Most likely a 0 (counterarguments ignored).</td>
<td></td>
</tr>
<tr>
<td>Constructive Politics</td>
<td>Most likely a 0 (positional politics).</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Participation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliberative</td>
<td>Most likely a 1.</td>
<td></td>
</tr>
<tr>
<td>Level of justification</td>
<td>Most likely a 2 (qualified justification) or a 3 (sophisticated justification).</td>
<td></td>
</tr>
<tr>
<td>Content of justification</td>
<td>Most likely a 0, 2 or 3 (less likely to be a neutral statement).</td>
<td></td>
</tr>
<tr>
<td>Respect</td>
<td>Most likely a 1 (implicit respect).</td>
<td></td>
</tr>
<tr>
<td>Counterarguments</td>
<td>Most likely a 1, 2, or 3 (less likely to ignore a counterargument).</td>
<td></td>
</tr>
<tr>
<td>Constructive Politics</td>
<td>Most likely a 1 (alternative proposal) or a 2 (mediating proposal).</td>
<td></td>
</tr>
</tbody>
</table>
7.2 Topic configuration

Moreover, we argue that it is not only deliberative quality that matters in a deliberative democracy, but the possibility of constituents to reach and debate with each other and their representatives. In this sense, the topic modelling conducted for the unlabelled data yields much more detailed information. The analysis does not produce a score, as in the DQI, but it provides a range of topics, categorized by thematic clusters and ranked by importance to the participants. In an age of big data and massive online participation, where it would be difficult to account for and summarize manually each contribution in a discussion of 25,000 comments, topic modelling offers an efficient and automated way to understand what was discussed. We therefore find that these thematic clusters, useful in feeding into the discourse analysis of any online political discussion. We detail here a list of political issues surfaced in the topic clusters for each dataset. As can be seen, even in discussions with a low Discourse Quality Category, such as the Republican Townhalls, a diverse range of relevant political issues were discussed.

<table>
<thead>
<tr>
<th>Political Issue</th>
<th>Datasets Discussed</th>
<th>Example n-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case processing center in rural Alberta – a big local employer</td>
<td>Trudeau Live Townhall</td>
<td>“help community deal”, “concerned community employed”, “case processing central”</td>
</tr>
<tr>
<td>Canada’s role in NAFTA</td>
<td>Trudeau Live Townhall</td>
<td>“good position renegotiate”, “chief economic adviser”</td>
</tr>
<tr>
<td>Canada’s Oversea Development Assistance (ODA) budget</td>
<td>Trudeau Live Townhall</td>
<td>“percent budget oda”, “international development institution”</td>
</tr>
<tr>
<td>Aboriginal issues</td>
<td>Trudeau Live Townhall, Trudeau Online Townhalls</td>
<td>“boil water advisory”, “need youth center”, “non indigenous Canadians”, “promise first nations”</td>
</tr>
<tr>
<td>Education reform</td>
<td>Trudeau Live Townhall, Republican Townhalls</td>
<td>“current education system”, “current standardized”</td>
</tr>
</tbody>
</table>
## Chapter 7. Discussion of results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Forum/Source</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brexit</td>
<td>EULive</td>
<td>“vote leave eu”, “leave leave leave”, “british londoners minority”</td>
</tr>
<tr>
<td>Immigration and globalization concerns</td>
<td>Trudeau Online Townhalls, Republican Townhalls</td>
<td>“save manufacturing jobs”, “put end globalization”, “blind immigration policies”, “taking muslim money”, “denounce muslim ban”</td>
</tr>
<tr>
<td>US social security</td>
<td>Bernie Sanders</td>
<td>“defend social security”, “need social security”, “worried social security”</td>
</tr>
<tr>
<td>Keystone and Dakota Access Pipelines</td>
<td>Republican Townhalls</td>
<td>“clean energy epa”, “support keystone pipeline”</td>
</tr>
<tr>
<td>Prisons closing</td>
<td>Republican Townhalls</td>
<td>“thoughts prisons closing”</td>
</tr>
<tr>
<td>LGBT rights</td>
<td>Republican Townhalls</td>
<td>“protect lgbt rights”</td>
</tr>
<tr>
<td>Mexican border wall</td>
<td>Republican Townhalls</td>
<td>“mexico pay wall”, “border wall important”</td>
</tr>
<tr>
<td>Women’s reproductive rights</td>
<td>Republican Townhalls</td>
<td>“money pay abortions”, “wall contraception women”, “affordable birth control”</td>
</tr>
<tr>
<td>US Health Care</td>
<td>Republican Townhalls</td>
<td>“single payer system”, “cover preexisting conditions”, “yes preventative care”</td>
</tr>
<tr>
<td>Electoral process</td>
<td>Republican Townhalls</td>
<td>“Pennsylvania election unconstitutional”, “efforts fix gerrymandering”</td>
</tr>
<tr>
<td>Climate change</td>
<td>Republican Townhalls</td>
<td>“combatting climate change”</td>
</tr>
<tr>
<td>Deliberative process</td>
<td>Republican Townhalls</td>
<td>“face town hall”, “real town hall”, “answer f*cking question”</td>
</tr>
</tbody>
</table>
Russian involvement in US elections

Republican Townhalls

“sanctions Russian ambassador”, “instructed Flynn mislead”, “hacking elections try”

Poverty reduction

UNDP Poverty Reduction

“innovation policy making”, precondition eradicating poverty”, “vulnerable therefore poor”, “declaration human rights”

In addition, we find an inverse correlation between cluster distribution and discourse quality. As we have seen in Chapter 4, it appears as though datasets with a less even distribution of documents in clusters also has higher individual rates of discourse quality, possibly because there was more convergence of issues in these discussions. An analysis of the correlation between the silhouette coefficient and the mean discourse quality give us a Pearson R score of -0.773, which means that they are negatively correlated (a score of 0 would be no correlation, and a score of 1 or -1 would mean maximal correlation).

Table 35: Silhouette coefficient and discourse quality category

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Silhouette coefficient</th>
<th>Mean Discourse Quality (0=low, 1=medium, 2=high)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trudeau Live Townhall</td>
<td>0.002</td>
<td>1.25</td>
</tr>
<tr>
<td>UNDP Poverty Reduction</td>
<td>0.024</td>
<td>0.924</td>
</tr>
<tr>
<td>Bernie Sanders</td>
<td>0.122</td>
<td>0.018</td>
</tr>
<tr>
<td>Republican Townhalls</td>
<td>0.130</td>
<td>0.013</td>
</tr>
<tr>
<td>Justin Trudeau Townhalls</td>
<td>0.141</td>
<td>0.031</td>
</tr>
<tr>
<td>EULive</td>
<td>0.308</td>
<td>0.013</td>
</tr>
</tbody>
</table>
7.3 Most important classifier features

During the instantiation phase, we used two classifiers to predict the deliberative quality of discursive comments using two sets of features: (a) bag-of-words and character count and (b) node-level. The following table shows the top features for each classifier used in each exercise.

Table 36: Most important features for determining discourse quality

<table>
<thead>
<tr>
<th>Mixed instantiation</th>
<th>Node instantiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook comment</td>
<td>Commenter frequency</td>
</tr>
<tr>
<td>Blog comment</td>
<td>Eccentricity</td>
</tr>
<tr>
<td>“behind”, “people”, “country”</td>
<td>Sentiment</td>
</tr>
<tr>
<td>Between 1500 and 3000 characters</td>
<td>Betweenness centrality</td>
</tr>
<tr>
<td>“one”</td>
<td>Degree centrality</td>
</tr>
<tr>
<td>Between 30 and 100 characters</td>
<td>Is leaf</td>
</tr>
<tr>
<td>Between 800 and 1500 characters</td>
<td></td>
</tr>
<tr>
<td>“also”, “must”, ”world”, “national”, “make”</td>
<td></td>
</tr>
</tbody>
</table>
### Chapter 7. Discussion of results

<table>
<thead>
<tr>
<th>less_than_30_characters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Live Town Hall</td>
<td></td>
</tr>
<tr>
<td>“let”, “years”, “countries”, “young”, “great”, “well”, “position”</td>
<td></td>
</tr>
</tbody>
</table>

In the following sections, we will discuss the relevance of character count, platform use (Facebook, blog or town hall) and word, along with the two top node features (commenter frequency and eccentricity).

#### 7.3.1 Character count

Recurring between 2 and 3 times in the top 10 features, character count is clearly an important feature when it comes to determining discourse quality. As we have seen, Blumenstock (2008) found that word count was the sole most important feature in determining the quality of Wikipedia articles. Their experiments with various classifiers using only one feature (over 2000 words) predicted quality with between 96% and 97% accuracy. On the other hand, n-gram bag of words classification produced only 81% accuracy. These findings appear to be very similar to the ones in our study.

The figure below shows the correlation between character count and DQI score.
Blumenstock (2008) argues that, in the case of Wikipedia articles, research measuring quality have often used complex analytical methods, whereas the article length measure is simple and accurate. In the case of online political discourse, one would argue that although it is not the only measure, the relationship between the two is also convincing.

7.3.2 Platform

Each classifier ranked “Facebook comment” as one of the top four features, while random forests and SVM used “Blog comment” and “Live Townhall” in second and fifth place, respectively. In this case, there are a few factors at play. Firstly, Fournier-Tombs (2013) found that political discourse via live chat was of significantly lower quality than political discourse as comments on a blog. The main reasons for this were character count and moderation. Character count, as addressed in the previous section, is a key platform differentiator.

The chart below shows character count per discussion platform, highlighting in particular the higher character count for blog comments.

Figure 24: Character count by DQI score in the training dataset
Chapter 7. Discussion of results

Secondly, in our study, blog comments had the additional feature of being moderated, usually by the original poster. There is a significant amount of research showing that moderation has a favourable impact on discourse quality (Diakopoulos and Naaman, 2011; Lampe et al, 2014). Park et al (2012), in their study on legislative rulemaking online in the United States, argue that facilitative moderation is a critical element in ensuring democratic participation. Similarly, in a study of participation in an online community of practice, Kilner and Hoadley (2005) find that moderation is key to conversations of higher deliberative quality.

7.3.3 Political words

The world list selected by the classifier is convincing in its relationship to political discourse. In the table below, we categorize the words according to seven categories – group, geographic, political, policy, argument, time and sentiment.

Table 37: Political words as classifier features

<table>
<thead>
<tr>
<th>Category</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group words</strong> – whose interests are being addressed?</td>
<td>“people”, “women”, “young”, “us”, “many”</td>
</tr>
<tr>
<td><strong>Geographic words</strong> – where are we situated in this discussion, at a local, community and national level?</td>
<td>“country”, “one”, “world”, “national”, “countries”, “communities”</td>
</tr>
</tbody>
</table>
Chapter 7. Discussion of results

<table>
<thead>
<tr>
<th>Political words – what political entities are we talking about, or to?</th>
<th>“party”, “leadership”, “left”, “convention”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy words – what issues are we addressing and how are we addressing them?</td>
<td>“policy”, “education”, “sustainable”, “eradicating”, “poverty”, “work”, “rural”, “goals”, “future”, “example”, “question”, “different”, “way”</td>
</tr>
<tr>
<td>Argument words – how are we making our point, how are we linking between one argument and another?</td>
<td>“also”, “must”, “make”, “need”, “let”, “believe”, “think”, “position”, “www”</td>
</tr>
<tr>
<td>Time words – what is the timeframe in which we are discussing?</td>
<td>“time”, “year”, “behind”, “ever”</td>
</tr>
<tr>
<td>Sentiment words – how are we expressing our feelings about this discussion?</td>
<td>“great”, “well”</td>
</tr>
</tbody>
</table>

7.3.4 Commenter frequency

Diakopoulos and Naaman (2011), in their study of comments on online newspaper articles, found that: "frequency of commenting is a valuable indicator of what moderators would consider low quality discourse". However, one might also expect to see more frequent commentors interacting more with their peers and thus obtaining higher scores in terms of counterarguments and constructive politics indicators. In our case, the correlation is not so obvious. The chart below shows comment count per commenter frequency score, by DQI group. Although comments with low DQI have a lower commenter frequency average than comments with high DQI, the trend is by no means clear. The chart below shows the distribution of comments by commenter frequency for each of the three DQI categories (in orange), along with a dashed blue line indicating the mean.
7.3.5 Eccentricity

Although this feature wasn’t prominent for the SVM classifier, it ranked highly for both the random forests and the logistic regression classifiers. Because the discussion network was structured as a tree, the measure of distance from root showed us how deep the comment was in the structure—answering other commenters rather than the original poster. This is important in several respects. First, there was a tendency for those who responded directly to the original poster to score highly in terms of content and level of justification, but much more poorly in terms of counterarguments and constructive politics. That is, they wrote complete and argumentative answers, but didn’t account for the answers of any of their peers. Those who responded to peers rather than the original poster were much more likely to take their arguments into account, whether to praise them or to disagree with them, and provide mediating proposals that would bring several viewpoints together.

However, like commentor frequency, eccentricity does not provide us with a definitive proxy for discourse quality. In fact, the low and medium-low DQI groups seem to have higher eccentricity than high and medium-high groups, trending towards lower DQI the further away from the root a node is.
Chapter 7. Discussion of results

7.4 Summary

In this chapter, we further discussed the results of the instantiation of the framework. We examined DQI category distribution, topic configuration and most important classifier features. The following chapter will conclude our study.
8. Conclusion

In this chapter, we return on the key elements of our study and outline its contributions, limitations, and a path forward for future research. We argued that the study of political discourse quality online is a current and important endeavor and that there are significant gaps in the development of automated discourse quality measurement methods. From this, we presented the DelibAnalysis framework, which provides a semi-automated means of evaluating discourse quality in a flexible, platform-agnostic way, and enhances it with a topic clustering method which enables the analyst to better understand the conversation. For this purpose, we used a random forests classification algorithm, along with a k-means clustering implementation using TF-IDF n-gram features. We discussed the technical implementation of the framework using Python, and provided code snippets as well as sample code in the Appendix. We then discussed design and algorithmic considerations that led us to propose our framework, before moving on to an instantiation of the framework using a mixed dataset of Facebook and blog data. Finally, we dove deeper into the analytical results of our DelibAnalysis instantiation and examined what features affect deliberative quality and what impact they might have on platform design choices in the future.

8.1 Contributions summary

As we have seen, the primary contributions of this study are a discourse quality evaluation framework that can be replicated across online political conversations, a technical implementation using Python, and an instantiation of the framework on Facebook and blog data which leads to insights into political quality online and design suggestions for the creation of platforms that enhance democracy. In this section, we will summarize these contributions and examine how they have answered the research questions presented in Chapter 1.

What factors determine whether an online political discussion is deliberative?

When we used the DelibAnalysis framework to evaluate our sample data, we extracted the most important features used in the random forests classification exercise. We also examined the features used in our experimental phase, which compared the performance of random forests to
logistic regression and SVM. We found that the most important factors in determining discourse quality in the online political discussions that we examined were character count and platform choice, and secondarily, commenter frequency and eccentricity.

**How can we automatically evaluate the deliberative quality of an online political discussion?**

In Chapter 4, we proposed the DelibAnalysis framework, a means of evaluating the deliberative quality of online discussion using a combination of classification (random forests) and clustering (k-means). We found random forests classification to be an appropriate way to generate a discourse quality category, which we found to have an up to 88% accuracy.

**What other means of understanding deliberative quality might be considered?**

In our DelibAnalysis framework, we added a k-means clustering topic analysis, which provides us with an overview of the main topic groups in a given conversation.

**8.2 Evaluation and limitations**

**8.2.1 Evaluation**

We used several approaches to evaluate the validity of our study. As we saw in Chapter 1, we began by evaluating the design of the research, specifically answering questions on construct validity and reliability of the research. We then evaluated the validity of framework components, by using accuracy scores for the classifier (the classification report and the f1-score), and the silhouette metric for the k-means clustering algorithm.

In order to evaluate the relevance of the framework as a whole, we used a demonstration approach, which aimed to show that it was possible to implement DelibAnalysis and that it did yield interesting results. We used a mixed dataset to show that the framework was platform-agnostic and that it would be possible to use it on a variety of datasets.
8.2.2 Limitations

Naturally, however, the work presented here was first limited by the type of data that was selected in order to test the framework. Although there was an attempt to select a broad range of data for analysis, it may be that other datasets would yield different factors influencing deliberative quality. As the research project is replicated across a variety of datasets, as we hope it will be in the future, we will begin to have a broader spectrum of possible factors.

In addition, this thesis was written at a time of fast-paced change in the area of machine learning. While cutting-edge a few years ago, the classification and clustering methods presented are beginning to be outpaced by new algorithms, particularly in deep learning. Although the advances in deep learning are currently limited to classification problems with extremely large training datasets and low interpretability, it is possible that in the future, a similar study run using deep learning for classification and clustering could yield better results.

8.3 Further usage

At this time, we hope that the framework presented here will be used by other researchers interested in exploring online discourse quality further. Additionally, we hope that the design suggestions presented here will be adopted in a pilot deliberation scenario that could eventually be adopted more broadly for democratic purposes.

8.4 Future work

In previous sections, we have alluded to several steps that would further the work undertaken here:

1. Re-develop the framework using a deep learning methodology and compare the results to what was obtained here, bearing in mind issues of interpretability;
2. Run the framework on a diverse set of data that would continue to inform us on the nature of online deliberation;
3. Pilot the design principles discussed in this study as an online deliberation platform.

As we saw in the introduction, current philosophical debates about democratic participation often turn towards the existential. Whether or not citizens have a real impact on the democratic process,
whether their voice is being heard – these questions take us back to the fundamentals of public sphere theory. When Habermas argued in the 1960s that technological changes were eroding the public sphere, he was referring to television and the way that passive watching provided citizens with information but with no means to react to the news in a participatory way. Now, however, the news is interactive. From the comfort of their homes, at work, on the subway, on vacation, citizens can not only receive information but they can correspond with their political representatives, react to legislative changes, make recommendations.

After this study, we conclude that structure, as in a blog, certainly adds to the democratic quality of a discussion. The objective is not to discourage citizens from dialoguing with their representatives on social media platforms such as Facebook, but to raise awareness to several key points. First, there is a distinction between voting democracy and participatory democracy, and attention must be given to the way in which the latter takes place in the online sphere. Second, the expectation that citizens may have that Facebook commenting is a tool for democratic participation should be tempered. While it is not out of the question that quality deliberation could take place on Facebook, this likely would not take place accidentally. That is, effort must be made to take into account the way in which quality deliberative conversations are structured online, using this as inspiration for deliberative design.

It is hoped that this research will contribute significantly in understanding the online public sphere and how it must be designed in order to promote participatory democracy.
9. Appendix A: Results from experimental project phase

*Table 38: Classification report using 4 DQI categories and random forests classifier and bag-of-words features*

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.68</td>
<td>0.85</td>
<td>0.76</td>
<td>55</td>
</tr>
<tr>
<td>Medium-low</td>
<td>0.43</td>
<td>0.08</td>
<td>0.13</td>
<td>39</td>
</tr>
<tr>
<td>Medium-high</td>
<td>0.25</td>
<td>0.64</td>
<td>0.36</td>
<td>14</td>
</tr>
<tr>
<td>High</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
</tr>
</tbody>
</table>

*Table 39: Prediction vs actual comparison using 4 DQI categories, random forests classifier and bag-of-words features*

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium-low</th>
<th>Predicted medium-high</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>47</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium-low</td>
<td>20</td>
<td>3</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium-high</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Actual high</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

As can be seen above, there was a strong bias towards scoring comments as low quality. Comments of high quality are not accurately scored at all. The highest F1-score, 0.76 for low quality, was still unsatisfactory.
We built a classifier using two different feature categories - character count and source of comment. The character counts were split into six separate features (less than 30 characters, between 30 and 100 characters, between 100 and 300 characters, between 300 and 800 characters, between 800 and 1500 characters, between 1500 and 3000 characters, more than 3000 characters). The source of comment were split into three separate features (facebook comment, live Townhall, blog comment).

This classifier was built with the same parameters as the bag of words classifier. Results were similar, as can be seen in the tables below.

Table 40: Classification report using 4 DQI categories, random forests classifier and quantitative features

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.82</td>
<td>0.84</td>
<td>0.83</td>
<td>55</td>
</tr>
<tr>
<td>Medium-low</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>39</td>
</tr>
<tr>
<td>Medium-high</td>
<td>0.26</td>
<td>0.93</td>
<td>0.41</td>
<td>14</td>
</tr>
<tr>
<td>High</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 41: Prediction vs actual comparison using 4 DQI categories, random forests classifier and quantitative features

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium-low</th>
<th>Predicted medium-high</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>46</td>
<td>1</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium-low</td>
<td>10</td>
<td>0</td>
<td>29</td>
<td>0</td>
</tr>
</tbody>
</table>
It was noticed that in both cases a significant number of medium-low comments were predicted medium-high. The two classifiers were therefore built again with a three category discourse quality index - low, medium and high.

The results are much better, as can be seen below.

9.1.1 Random Forest classification using bag-of-words only

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.87</td>
<td>0.96</td>
<td>0.92</td>
<td>84</td>
</tr>
<tr>
<td>Medium</td>
<td>0.71</td>
<td>0.62</td>
<td>0.60</td>
<td>23</td>
</tr>
<tr>
<td>High</td>
<td>1.00</td>
<td>0.40</td>
<td>0.57</td>
<td>5</td>
</tr>
<tr>
<td>Average/total</td>
<td>0.84</td>
<td>0.85</td>
<td>0.84</td>
<td>112</td>
</tr>
</tbody>
</table>

Table 43: Prediction vs actual comparison using 3 DQI categories, random forests classifier and bag-of-words features

| Actual medium-high | 0 | 1 | 13 | 0 |
| Actual high        | 0 | 4 | 0  | 0 |
9.1.2 Random Forests classification using character count and comment category only

Table 44: Classification report using 3 DQI categories, random forests classifier and quantitative features

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.93</td>
<td>0.88</td>
<td>0.90</td>
<td>84</td>
</tr>
<tr>
<td>Medium</td>
<td>0.61</td>
<td>0.74</td>
<td>0.67</td>
<td>23</td>
</tr>
<tr>
<td>High</td>
<td>1.00</td>
<td>0.80</td>
<td>0.89</td>
<td>5</td>
</tr>
<tr>
<td>Average/total</td>
<td>0.87</td>
<td>0.84</td>
<td>0.85</td>
<td>112</td>
</tr>
</tbody>
</table>

Table 45: Prediction vs actual comparison using 3 DQI categories, random forests classifier and quantitative features

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium-low</th>
<th>Predicted medium-high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>74</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium</td>
<td>6</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Actual high</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

9.1.3 SVM classification

The results of the SVM classification were distributed as follows:
### Table 46: Classification report using 3 DQI categories, SVM classifier and hybrid features

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.84</td>
<td>0.87</td>
<td>0.85</td>
<td>84</td>
</tr>
<tr>
<td>Medium</td>
<td>0.42</td>
<td>0.43</td>
<td>0.43</td>
<td>23</td>
</tr>
<tr>
<td>High</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>5</td>
</tr>
<tr>
<td>Average/Total</td>
<td>0.71</td>
<td>0.74</td>
<td>0.73</td>
<td>112</td>
</tr>
</tbody>
</table>

### Table 47: Predicted vs actual report using 3 DQI categories, SVM classifier and hybrid features

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>73</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium</td>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Actual high</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
9.1.4 Logistic Regression classifier

The results of the Logistic Regression classification were distributed as follows:

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.85</td>
<td>0.88</td>
<td>0.87</td>
<td>84</td>
</tr>
<tr>
<td>Medium</td>
<td>0.50</td>
<td>0.52</td>
<td>0.51</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 48: Classification report using 3 DQI categories, logistic regression classifier and hybrid features
### Table 49: Predicted vs actual report using 3 DQI categories, logistic regression classifier and hybrid features

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>74</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium</td>
<td>11</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Actual high</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
9.1.5 SVM classification using node-level features

Table 50: Classification report using SVM classifier and node-level features

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>Medium-low</td>
<td>0.67</td>
<td>0.29</td>
<td>0.40</td>
<td>7</td>
</tr>
</tbody>
</table>
### Table 51: Prediction vs actual report using SVM classifier and node-level features

<table>
<thead>
<tr>
<th></th>
<th>Predicted low</th>
<th>Predicted medium-low</th>
<th>Predicted medium-high</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium-low</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium-high</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Actual high</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
9.1.6 Logistic regression classification using node-level features

Table 52: Classification report using logistic regression classifier and node-level features

<table>
<thead>
<tr>
<th>Discourse Quality</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Medium-low</td>
<td>0.58</td>
<td>1.00</td>
<td>0.74</td>
<td>7</td>
</tr>
<tr>
<td>Medium-high</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>5</td>
</tr>
<tr>
<td>High</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>Average/Total</td>
<td>0.34</td>
<td>0.58</td>
<td>0.43</td>
<td>12</td>
</tr>
</tbody>
</table>
Table 53: Predicted vs actual report using logistic regression classifier and node-level features

<table>
<thead>
<tr>
<th>Actual level</th>
<th>Predicted low</th>
<th>Predicted medium-low</th>
<th>Predicted medium-high</th>
<th>Predicted high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Actual medium-low</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Actual high</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The results of the predictions from each classifier were 56% accuracy for the random forests classifier, 57% accuracy for the SVM classifier and 34% accuracy for the logistic regression classifier. We can conclude the both random forests and SVM are performing in a similar manner but that logistic regression, in this case, did not yield satisfactory results. However, the overall accuracy of the classifiers is low, especially when one compares with the 88% accuracy of random forests using hybrid features. At the moment, this is attributed to the fact that a small training dataset is used. Additional data will be trained in order to attempts to improve the results.
10. Appendix B: Comment Examples

From Section 4.3: “I’d very much like to comment on something that has been less discussed hitherto. We have dealt with EDUCATION, which of course is vital, and with MEASUREMENT and DATA, without which we’d be navigating without a compass. However, the other essential element of SD has not been that much present. Three days ago Carmen Ghanim has dealt with the ENVIRONMENT. And yes, Carmen, how right you are! Of course, you may tell me that without EDUCATION there can be no concern as far as the ENVIRONMENT is concerned; right you are too. Sustainable Development must be tackled from its three axes. Carmen says: "What can we do [about it]?? -Start building back what has been destroyed: THE MOST IMPORTANT STEP is that of re-planting Forests WORLD-WIDE!!" In our fertile Argentine country-side the last decade saw the destruction of native and planted forest areas in order to grow soy bean. Soy bean has been in great demand in order to be exported to China (it has nothing to do with feeding the local population nor our cattle). This destruction of forest areas is responsible for DESERTIFICATION. Desert areas, where forest stood, means what all deserts mean, plus huge clouds of dust generated where forests stood. Today the indiscriminate sowing of soy bean is being balanced by means of wheat crops and areas where cattle grow to produce the unique meat brand Argentina has been famous for. I’m far from being an expert in this field, so I ask my colleagues, especially my Argentine colleagues who may read this, to bear with me. But it is clear that without stopping forest destruction and switching to re-planting, there is NO FUTURE. For you all know that forests absorb part of the huge amount of tons of greenhouse gas emissions all our countries are generating every day. Of course, the oceans still absorb greenhouse gas emissions; but we all know how deeply the oceans have been affected! And to tackle these combined tasks efficiently, apart from an unrelenting struggle against corruption, EDUCATION must also point towards re-educating the government officials responsible, and those officials who are struggling towards leaving no one behind.”
11. Appendix C: DelibAnalysis Codebase

**DelibAnalysis Classification**

The following script implements the random forests classifier in order to predict the Discourse Quality Index (DQI) category of online comments.

```python
from sklearn.metrics import classification_report
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
import nltk
import re
from nltk.corpus import stopwords
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('ggplot')
```
In [ ]: # Functions to process data
def process_labelled_data(source_csv):
    data_df = pd.read_csv(source_csv)
    indicators = ['participation', 'level_of_justification', 'content_of_justification', 'respect', 'counterarguments', 'constructive_politics']
    data_df['dq_i'] = data_df[indicators].sum(axis=1)
    data_df['dq_i_groups'] = data_df.dqi.map(lambda x: 0 if x <= 5 else
    1 if (x > 5 and x <=10) else 2)
    data = data_df[['dq_i', 'comment', 'dq_i_groups','fb_comment', 'live_t h', 'blog_comment']]  
    return data

def comment_to_words(raw_comment):
    try:
        letters_only = re.sub(r'[^a-zA-Z]', '', raw_comment)
        words = letters_only.lower().split()
        stops = set(stopwords.words('english'))
        meaningful_words = [w for w in words if not w in stops]
        return(' '.join(meaningful_words))
    except TypeError:
        print raw_comment

def append_features(input_matrix,input_feature):
    count = 0
    new_matrix = np.zeros(shape=(input_matrix.shape[0], input_matrix.shape[1]+1))
    for i in range(0, len(input_feature)):
        new_matrix[:,i] = np.append(input_matrix[:,i], input_feature[i])
    return new_matrix

def add_character_counts(data, chars):
    data['char_count'] = data['comment'].apply(lambda x: len(x))
    for k, v in chars.items():
        data[k] = data.char_count.map(lambda x: 1 if (x <= v[0] and x > verr[1]) else 0)
    return data

In [ ]: # Upload and process labelled data
labelled_data = process_labelled_data("combined_scored.csv")
labelled_data["cleaned_comment"] = labelled_data["comment"].apply(lambda x: comment_to_words(x))
labelled_data = add_character_counts(labelled_data, char_dict)
print(labelled_data.head())
In [ ]:

```python
# Train classifier

train, test = train_test_split(labelled_data, train_size = 0.8, random_state = 44)

vectorizer = CountVectorizer(analyzer = "word", tokenizer = None, preprocessor = None, stop_words = None, \       max_features = 6000)

train_data_features = vectorizer.fit_transform(train["cleaned_comment"])

train_data_features = train_data_features.toarray()

print train_data_features.shape

quantitative_features = ["fb_comment", "live_th", "blog_comment", "less_than_30_chars", \      "between_30_and_100_chars", \      "between_100_and_300_chars", "between_300_and_800_chars", \      "between_800_and_1500_chars", \      "between_1500_and_3000_chars", "more_than_3000_chars"]

for i in quantitative_features:
    train_data_features = append_features(train_data_features, train[i].as_matrix())

print '(Number of comments, number of features)'
print train_data_features.shape
```

In [ ]:

```python
# Create the classifier

forest = RandomForestClassifier(n_jobs=-1, n_estimators=24, criterion="entropy", max_depth=17, warm_start=True, \      max_features=2000, bootstrap=True)

y, _ = pd.factorize(train["dqi_groups"])

forest.fit(train_data_features, y)
```
Appendices

```python
In [ ]:
# View the top features used by the classifier by importance
importances = forest.feature_importances_
std = np.std([tree.feature_importances_ for tree in forest.estimators_], axis=0)
indices = np.argsort(importances)[::-1]
vocab = vectorizer.get_feature_names()
for i in quantitative_features:
    vocab.append(i)

# Print the feature ranking
print("Feature ranking:")
feature_importance_df = pd.DataFrame(data=None, columns=['Feature name', 'Importance'])
for f in range(0, 50):
    feature_importance_df.loc[f+1] = [vocab[indices[f]], importances[indices[f]]]
plt = feature_importance_df.plot(kind="barh", figsize=(10,10), color="purple")
plt.set_yticklabels(feature_importance_df["Feature name"])
plt.invert_yaxis()
plt.set_title("Top 50 features by importance")
plt.set_xlabel("Importance")
plt.set_ylabel("Feature name")
```

```python
In [ ]:
# Process the test data
test_data_features = vectorizer.transform(test["cleaned_comment"])
test_data_features = test_data_features.toarray()
print test_data_features.shape
for i in quantitative_features:
    test_data_features = append_features(test_data_features, test[i].as_matrix())
print '(Number of comments, number of features)'
print test_data_features.shape
```

```python
In [ ]:
# Evaluate the classifier by predicting the score of the test group
result = forest.predict(test_data_features)
output = pd.DataFrame(data={"actual_dqi": test["dqi_groups"], "predicted_dqi": result})

# Create confusion matrix
print pd.crosstab(output['actual_dqi'], output['predicted_dqi'], rowname s=['Actual'], 
        colnames=['Predicted'])
print '
Classification Report:
', classification_report(output['actual_dqi'], output['predicted_dqi'])
```
DelibAnalysis Clustering

The following script implements the k-means clustering algorithm in order to understand the topics of online comments.

```python
# Import libraries
from sklearn.datasets import fetch_20newsgroups
from sklearn.decomposition import TruncatedSVD
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import HashingVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import Normalizer
from sklearn import metrics

%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('ggplot')

from sklearn.cluster import KMeans, MiniBatchKMeans

import sys
from time import time
import pandas as pd
import re
from nltk.corpus import stopwords
import numpy as np

# Clean dataset
dataset = pd.read_csv('combined_scored.csv')
labels = dataset.columns

vectorizer = TfidfVectorizer(use_idf=True, max_features=1000, analyzer='word', ngram_range=(2,3))

def comment_to_words(raw_comment):
    letters_only = re.sub(r'[^a-zA-Z]', '', raw_comment)
    words = letters_only.lower().split()
    stops = set(stopwords.words('english'))
    meaningful_words = [w for w in words if not w in stops]
    return ' '.join(meaningful_words)

clean_train_comments = []

dataset['cleaned_comment'] = dataset['comment'].apply(lambda x: comment_to_words(x))

fit_vectorizer = vectorizer.fit_transform(dataset['cleaned_comment'])
svd = TruncatedSVD(n_components=100)
normalizer = Normalizer(copy=False)
lsa = make_pipeline(svd, normalizer)
fit_lsa = lsa.fit_transform(fit_vectorizer)
```
In [ ]:

```python
# Implement KMeans

km = KMeans(n_clusters=5, init='k-means++', max_iter=100, n_init=10,
            verbose=True)
km.fit(fit_lsa)

print("Silhouette Coefficient: 10.36")
print(metrics.silhouette_score(fit_lsa, km.labels_, sample_size=1000))

print("Top terms per cluster:")

original_space_centroids = svd.inverse_transform(km.cluster_centers_)
order_centroids = original_space_centroids.argsort()[::-1]
terms = vectorizer.get_feature_names()
for i in range(5):
    print("Cluster %d:" % i, end='')
    for ind in order_centroids[i, ::10]:
        print('; is ' % terms[ind], end='')
    print()
```

In [ ]:

```python
# Visualise top 15 3-grams by importance

t = vectorizer.idf_.
idf_dict = dict(zip(vectorizer.get_feature_names(), idf_))
idf_df = pd.DataFrame.from_dict(idf_dict, orient='index')
idf_df = idf_df.sort_values(by=0, ascending=False)
for i in range(5):
    df = pd.DataFrame(columns = ['ngram', 'tf-idf-score'])
    for ind in order_centroids[i, :15]:
        df2 = pd.DataFrame([[terms[ind], idf_df.get_value(index=terms[ind], col=0)]], columns = ['ngram', 'tf-idf-score'])
        df = df.append(df2)
    df = df.sort_values(by='tf-idf-score', ascending=False)
    plt = df.plot(kind='barh', legend=None, color='purple')
    plt.invert_yaxis()
    plt.set_yticklabels(df['ngram'])
    plt.set_title("Top 15 3-grams by importance: Cluster " + str(i))
    plt.set_xlabel("TFIDF Score (importance in dataset")
    plt.set_ylabel("3-grams")
```
12. References


Chapter 11. References


Finch-Lees, Tess. "Brexit is not the high point of democracy – it’s the greatest fraud ever perpetrated in British politics". *The Independent* February 2, 2017 (2017)


Chapter 11. References


Freelon, Deen G. "Analyzing online political discussion using three models of democratic communication." New Media & Society 12.7 (2010): 1172-1190.


Oremus, Will. "How many people really get their news from Facebook?". *Slate* December 20, 2016 (2016).


Small, Mario Luis. Weak ties and the core discussion network: Why people regularly discuss important matters with unimportant alters, *Social Networks*, 35.3 (2013): 470-483


