



Alkharashi, Abdulwhab A. (2021) *Exploring the characteristics of abusive behaviour in online social media settings*. PhD thesis.

<https://theses.gla.ac.uk/82389/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

EXPLORING THE CHARACTERISTICS OF ABUSIVE BEHAVIOUR IN ONLINE SOCIAL MEDIA SETTINGS

ABDULWHAB A. ALKHARASHI

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING SCIENCE
COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

AUGUST 16, 2021

© ABDULWHAB A. ALKHARASHI

Abstract

Online abusive behaviour can impact interaction amongst contributors and moderators. It may lead to physical harm or threats. Existing research has not addressed the perception of moderation activity, discussion and disagreement can cause contributors to react aggressively.

This thesis investigates the factors that lead to abusive behaviour in conversations within online settings. In particular, empirical analyses were conducted to identify the factors that contribute to abuse in online settings and to distinguish between polite and abusive forms of disagreement. Three contributions were presented in this research to address each to social computing, computational social science and cyber abuse research domains.

The analyses suggested that moderators on Reddit view themselves as members of their community and work hard to both guard against violations, but also with contributors to enhance the quality of their content. Moderators also reported the nuances that distinguish polite and abusive disagreement.

Furthermore, the analyses revealed that the differences between in-person and online conversations can help identify abusive behaviour. Specifically, the setting of discussion fosters participant behaviours (less hedging, more extreme sentiment, greater willingness to express personal opinion and straying from topic) that are known to increase the likelihood of abusive behaviour. Additionally, the findings revealed how consensus-building factors can influence disagreement in different settings.

Finally, we showed how disagreement can be identified and can affect votes based on linguistics contexts. It was shown that different forms of disagreement can be detected better when using specific abuse, politeness and sentiment textual features using models of multi-label text classification.

The above research findings conceptualised the development of moderation systems to combat online abusive behaviour, based on analysis of the type of disagreement a contribution embodies and other linguistic and behavioural characteristics.



Acknowledgements

"Praise be to God, whose glory and majesty will complete good works, O Lord to you Praise as it should for the majesty of your face and the greatness of your authority. And the punishment of the hereafter, God, whoever shows the beautiful and the ugly cover, who does not take offence and does not violate the concealment, great pardon and good transgression".

First and for the most, I would like to thank my primary supervisor Dr Tim Storer for the endless support– without his support this work wouldn't be completed. I extend my gratitude to the supervisory team: Prof Andrew Hoskins and Prof Joemon Jose for their insightful comments and valuable meetings. I also would like to thank the undergraduate students from the School of Computing Science at the University of Glasgow and the moderators of Reddit platform for their valuable time and participating. I am grateful to the Saudi Electronic University for funding my research. I would like to acknowledge Human Data Interaction (EPSRC Network Plus) for partially funding this research (Grant Number EP/R045178/1) led by Dr Catherine Happer.

All of my colleagues at the school were very friendly and supportive including my office mates: Mohammed Alhamad, Saad Altamimi, Gibrail Islam and Tom Wallis. It was a sad news to learn about the lost of our former PhD student Robbie Simpson. I would like to thank Ibrahim Alghamdi, Saad Alahmari and Abdullah Almayouf for their companionship. I extend my appreciation to Mohamed Khamis for his time and perspicacious discussions. I would like to thank John the Janitor for morning conversations about Scotland ancient history and wisdom.

Last but not least, I would like to thank my sincere mother for her daily encouragement and check ups. I also, would like to thank my wonderful wife for her support during the difficult period of the PhD. Of course, my son 'Abdullah' for keeping me happy whenever I get back home after research work.

Dedication. (This thesis is dedicated to my parents)

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Table of Contents

Abstract

Acknowledgments

1	Introduction	1
1.1	Background	1
1.2	Motivation	2
1.3	Thesis Statement and Research Questions	4
1.4	Contributions	6
1.4.1	Understanding perceptions on content moderation	6
1.4.2	Online vs. in-person behaviour in conversations	7
1.4.3	Disagreement and abuse in online discussions	8
1.5	Thesis Overview	9
1.6	Publications	10
2	Literature review	11
2.1	Introduction	11
2.2	Definitions and forms of abusive behaviour	12
2.3	Moderators and their view point	16
2.3.1	Moderation impact on contributors	16

2.3.2	Behaviours and perceptions correlated with moderation	20
2.3.3	Research gap	22
2.4	Causes of abusive behaviour online	22
2.4.1	Community reactions	23
2.4.2	Situational factors	24
2.4.3	Anonymity	25
2.4.4	Online vs in-person	25
2.4.5	Social norms	27
2.4.6	Research gap	28
2.5	Methods for tracking abusive content, performance of automated tools	29
2.5.1	Contextual-based detection	29
2.5.2	Activity-based detection	33
2.5.3	Assistive approaches	35
2.5.4	Research gap	36
2.6	Summary	36
3	Moderators vs Contributors: The Case of Perspectives on Content Mod- eration	38
3.1	Introduction	38
3.2	Forms of moderation	41
3.3	The Reddit platform	42
3.4	Method	43
3.4.1	Survey of Moderators	44
3.4.2	Collection of Moderator Action Data	45
3.5	Findings	46
3.5.1	Demographics	47
3.5.2	Moderator roles	47

3.5.3	Working patterns	49
3.5.4	Intervention actions	52
3.5.5	Reasons for intervention	54
3.5.6	Topics for intervention	56
3.6	Discussion	57
3.6.1	Implications for the development of moderation system	57
3.6.2	Limitations	58
3.7	Summary	59

4 Online vs In-person Conversation: The Case of a Peer-Group Project in a Learning Environment 61

4.1	Introduction	61
4.2	Related Work	65
4.2.1	Online disinhibition	65
4.2.2	Peer-group interaction in online and in-person discussion	66
4.2.3	Measuring online and in-person behaviour	67
4.3	Case Study	68
4.4	Experimental Design	68
4.4.1	Overview	68
4.4.2	Course structure	70
4.4.3	Recruitment strategy	71
4.4.4	Pre-discussion (Who?)	71
4.4.5	Meeting iterations	72
4.4.6	Post-discussion (Why?)	72
4.4.7	Ethics	73
4.4.8	Dependent variables	74
4.5	Results	75

4.5.1	Participants	75
4.5.2	Data set	76
4.6	Measures of Polite and Abusive Text (Q3)	77
4.6.1	Experience and background	77
4.6.2	Conversations	77
4.7	Understanding and Identifying Stimulated Behaviour (Q4)	81
4.7.1	Evaluation of participants	81
4.7.2	Causality between features	86
4.7.3	Classifying abusiveness based on features	88
4.8	Qualitative Analysis	91
4.8.1	Method	91
4.8.2	Consensus building factors	92
4.9	Discussion	95
4.9.1	Technical design implications	95
4.9.2	Limitations	96
4.10	Summary	96
5	Polite vs Abusive Disagreement: The Case of Polemicists	98
5.1	Introduction	98
5.2	Background	101
5.2.1	Disagreement levels	101
5.2.2	Disagreement detection	103
5.3	Overview of Methodology	104
5.3.1	Preparation procedure ①	105
5.3.2	Crowdsourcing set-up ②	107
5.3.3	Classification ③	110

5.3.4	Ethics	111
5.4	Understanding Factors of Disagreement and Identifying Abuse	111
5.4.1	Factors of disagreement	112
5.4.2	Vote abuse	117
5.4.3	Capturing disagreement	118
5.4.4	Predictive analysis	121
5.5	Discussion	124
5.5.1	Current design implications	124
5.5.2	Theoretical implications	125
5.5.3	Limitations	125
5.6	Summary	127
6	The Interplay between Disagreement, Abuse and Moderation in Online Discussions	129
6.1	Introduction	129
6.2	Moderation anticipatory system	131
6.3	Online and in-person differences and similarities	134
6.4	Detecting abuse and disagreement	136
6.5	Summary	139
7	Conclusion	142
7.1	Review Thesis Statement and Research Questions	143
7.1.1	RQ1 and RQ2	143
7.1.2	RQ3 and RQ4	144
7.1.3	RQ5	145
7.2	Future Directions	146
7.2.1	Possible effects of abusive behaviour in discussions	146
7.3	Concluding Remark	148

Bibliography **148**

A An Appendix **176**

 A.1 Pre-discussion questionnaire 176

 A.2 Post-discussion questionnaire 177

 A.3 Demographics of participants 178

List of Tables

1.1	Summary of experiments by RQs	6
2.1	Definitions of online abusive behaviour	13
2.2	Performance for the proposed automated text detection methods	30
2.3	Textual features of identifying abusive behaviour	33
3.1	The survey questions for moderation themes	45
3.2	Three-dimensional background factors	53
4.1	Summary of participants for each demographic group.	75
4.2	Descriptive summary of conversations dataset	77
4.3	Statistical significance between online and in-person groups	81
4.4	Post-discussion responses	82
4.5	Post-discussion responses (Cont.)	83
4.6	Post-discussion responses (Cont.)	84
4.7	Classifiers performance for predicting settings and removed comments	90
5.1	Summary of the collected data from the reddit	105
5.2	Task assignment of each round per batch	110
5.3	Example comments by disagreement level	118
5.4	Classification task performance for detecting disagreement	119

A.1 Demography of online and in-person participants 179

List of Figures

3.1	Summary of demographics ($N = 218$) recruited from reddit	47
3.2	Moderator views on their roles.	48
3.3	Time commitment and reviews	50
3.4	Hourly activity of moderators vs contributors on Reddit	51
3.5	Actions taken to intervene on posts.	52
3.6	Categories of reasons for interventions	55
3.7	Categories of topics for interventions	57
4.1	Overview of steps for the first part of quantitative analysis.	69
4.2	Process for conducting the experimental study.	70
4.3	Example of online/in-person group discussion	73
4.4	Word cloud for online and and in-person conversations	76
4.5	Frequency responses of pre-survey (Q7)(Q8)(Q12)	78
4.6	Familiarity responses of pre-survey (Q10)(Q13)(Q14)	78
4.7	Textual factors among online and in-person discussions	80
4.8	Examples of comments associated with polite and abuse textual features	86
4.9	Graphical mode of Bayesian networks that shows the conditional probability between each investigated variable. For example, there is high chance of probability (90%) that negative emotion is leading to profanity in the same comment.	87

4.10	A flow diagram of text classification tasks	89
4.11	Mind map of most cross-frequent phrases	92
5.1	An example of disagreement and vote abuse	99
5.2	The levels of disagreement scale	102
5.3	Steps for building disagreement classifiers	104
5.4	Samples of top three significant textual features	108
5.5	An example of HIT task with instructions	109
5.6	Precautionary steps in crowdsourcing experiment.	111
5.7	Disagreement on politeness factors	114
5.8	Disagreement on abuse factors	114
5.9	Disagreement on abuse factors	115
5.10	Response time of disagreement levels	116
5.11	Summary of the three experiments, multi-label and multi-class classification tasks.	120
5.12	Bayesian networks model for post- and pre-disagreement	121
5.13	Predictive analysis for disagreement vs vote and duration	123
6.1	Summary of the thesis contributions	130

Chapter 1

Introduction

"One of the biggest challenges will be finding an appropriate balance between protecting anonymity and enforcing consequences for the abusive behaviour that has been allowed to characterize online discussions for far too long." Poland [1]

This chapter provides an overview of the research background, motivation, thesis statement, research questions, contribution and structure of this thesis. The chapter is comprised of six sections. Section 1.1 presents the research problem; it is followed by the motivation of this research in Section 1.2. Section 1.3 outlines the thesis statement, and presents the research questions and experiment outline. Section 1.4 describes the key contributions of this research. Section 1.5 highlights the thesis structure of each chapter.

1.1 Background

Social media sites have become a necessary method of interaction to share news or post daily activity; this user-generated content includes text and media files. For instance, Reddit is an aggregated news and social platform that hosts massive online communities and users. The number of average monthly active users in Reddit is in excess of 430 million with more than 30 billion views monthly, hosting at least 130 active

online communities [2]. There are also online community-based platforms that offer different features of communication, e.g., Facebook, Google+, Twitter, Snapchat, Tik Tok and so on. However, these social platforms tend either rely on auto-moderation or users' flagging or reporting for violating the rules.

Multiple forms of reported online abuse include trolling (posting disruptive comments to destruct other users), cyberbullying (online social aggression to intimidate individuals), swatting (making false calls to target victim online) behaviours, etc., can alienate users from a particular community [3, 4]. Gillespie [5] interviewed active moderators, community designers and contributors to understand the motives for content removal and other actions against trolls and cyberbullies as a key domain of public and political discourse, and concluded that current approaches for content moderation have limited effectiveness.

Similarly, Roberts [6] conducted interviews with commercial content moderators from the west coast of the US and the Philippines to reveal the identity of moderators and investigate the obstacles that interplay the decision of human interventions. The authors found that moderators mostly hide the fact that they moderate communities and receive comparatively low wages for doing such job.

Abusive behaviour in online forums can be disruptive to the focus and direction of a discussion [7]. In some cases, this may go further, causing disruption to the community [8] to the detriment of other participants, other social harms [9] or in extreme cases direct personal related to gender abuse [10] or even physical harms [11]. Many forums employ moderators to regulate discussions, however, the scale of social media may well overwhelm their ability to control discussions, particularly during periods of intense activity [12].

1.2 Motivation

Despite the popularity and scalability of online discussion communities or social networks sites, online community users continue to suffer from abusive behaviour in all kinds. The users join variety of communities to express strong opinions against or for political figures, social events or religious beliefs. These arguments can often encour-

age abusers to practise forms of hacktivism [13, 14, 15], acts that intend to breach users' identity and data without permission for political or personal reasons or to promote any illegal activity.

Suler [16] proposed six factors that disinhibit online communication as a result of the anonymity of the internet. However, research around the causes and forms of abusive behaviours in online communities remains scant and at the preliminary stage of exploration. Wilson and Kelling [17] introduced the Broken Windows Theory which suggests that if local authorities cannot control actions of abusive behaviour from community members, people are more likely to come back and do the same thing. This will increase the likelihood of apathy and risk harming civilians. This also applies to online community members and moderators.

Prior work studied the social phenomenon of online abuse within qualitative methods surveys and interviews that investigate the motives, actions and reactions by self-reporting the experiences or expectations of such behaviour online. For example, Buckels et al. [18] invited participants from Amazon Mechanical Turk to complete an online survey about the way they post comments online. The authors then conclude that there is a strong correlation between sadism and abusive behaviour, and the frequency of posting. The sadism in this context refers to taking pleasure from upsetting others. Another study [19] used interviews with Wikipedians to find more about motives of abusive behaviour and reported that abusers took advantage of the anonymity of identity to disrupt the productivity of knowledge-based communities.

Numerous methods were proposed to combat or reduce the aggressively abusive behaviour in online communities including: text classification [20, 21], deep learning [22, 23, 24], leveraging crowd-sourcing [25, 26], characterisations and activities of users [27, 28], moderation tools and approaches [29, 30], community feedback [31, 32].

A recent study by Pew Research centre [33] reports that at least one in five people experience forms of online harassment, especially young females. Additionally, Smith and Duggan [34] found that disagreement can lead to severe in-person death threats.

Abusive behaviour is defined and constructed normally by the community guidelines and expectations. Most of these platforms rely on conventional methods to discourage undesirable behaviour, e.g., moderation or report posts, votes, mute posts, and entirely banning users' ability to post. Abusive behaviour on any online platform can lead to

major risks and concerns outside the community including violence, harassment and threats [16]. Social media is generally considered to be a more fertile ground for abusive behaviour, yet the causes of this are not well understood. The development of moderator practices has largely been ad hoc and unstudied. We do not know how moderators should best react to different forms of undesirable behaviour in different contexts. We do not really understand how legitimate disagreement can deteriorate into undesirable abusive behaviour, or how to detect whether this is happening on multiple scales. Moderators can therefore be overwhelmed. Moderators of online discussions also respond differently, engaging in specific tactics to maintain debate quality that are specific to settings. In particular, what seems to be acceptable in one community may not be acceptable in another. Current work is interested in understanding the cases of abusive content and how they can interplay the online contributions overall.

1.3 Thesis Statement and Research Questions

This thesis asserts that:

Contributions to online discussions can be detected by classifying contributions in terms of the form of disagreement that they embody.

Abusive behaviour is contextual and may be conflated by community participants, with disagreeable or controversial contributions (e.g. through down-votes), exacerbating the workload of community moderators. Further, online behaviour can be shown to be quantitatively more prone to disagreement and abusive behaviour than in-person, due to the lack of wider social cues and ‘guard rails’. Finally, these insights allow us to classify behaviour in terms of the form of disagreement, distinguishing between polite and abusive disagreement.

Therefore, this thesis investigates the extent to which people disagree differently with one another online, compared to in-person. Investigating this, together with a means of understanding different forms of disagreement online is essential to understand the mechanism of communication in different settings to support the design of better moderation systems.

To investigate the above statement, this thesis seeks to answer the underlying questions summarised by chapter and experiment in Table 1.1 :

- RQ1. What role do moderators perceive for themselves on Reddit discussions?
- RQ2. When, how and why do moderators intervene on discussions on Reddit?
- RQ3. Is there a statistically significant difference between online and in-person discussions in terms of polite or abusive language used? Can conversation settings be detected?
- RQ4. To what extent can stimulated behaviour shape the understanding and perceptions of peer-group evaluation and consensus in discussions?
- RQ5. What kind of context enables and promotes polite or abusive disagreement on an online discussion? Do particular kinds of disagreement trigger down voting?

To investigate these questions, moderators were surveyed to understand the nature of online disagreement. In addition, data sets were collected from different types of conversations both online and in-person, and at scale from the Reddit online social media platform in order to evaluate the differences in disagreement of interest.

Reddit is an appropriate target to investigate due to the diversity of its discussion forms (more than 2 million communities/subreddits and 430 monthly active users) [35], and extensive moderator practice. The platform can help researchers to conduct studies of computer human interactions (HCI), and social computing, and computational social science. For example, norms violations [36], engagement between users and moderators [37] and attributes of rules [38]. Additionally, programming libraries are available for data collection and analysis, i.e., Python Reddit API Wrapper ¹ , Reddit API ² and pushshift ³

¹<https://praw.readthedocs.io>

²<https://www.reddit.com/dev/api>

³<https://pushshift.io>

Experiment	Chapter	Research Questions	Purpose
Experiment 1	CH3	RQ1 RQ2	<ul style="list-style-type: none"> • Understand roles and perceptions of moderators • Learn about intervention issues and strategies
Experiment 2	CH4	RQ3 RQ4	<ul style="list-style-type: none"> • Differences between online and in-person discussion • Understand factors of stimulated and consensus behaviours
Experiment 3	CH5	RQ5	<ul style="list-style-type: none"> • Present classifications & analyses of polite and abusive disagreement

Table 1.1: Summary of experiments for the listed research questions aimed to answer in this thesis

1.4 Contributions

This thesis addresses the above research questions and makes distinct contributions to the two fields of computational social science and social computing research domains. In particular, Computational social science is the application of statistical methods, machine learning, social network analysis and other computational techniques within social science research. The techniques are used to address questions relating to human behaviour within societies and larger scale social phenomena [39]. Social computing research focuses on the design and evaluation of computing platforms that support and enable social interactions. This includes systems that enable communications between humans and between computers and humans, as well as guidelines and tools that support the mediation and moderation of communication [40].

The following subsections explain the contributions of the thesis within these two specific contexts. In addition, the thesis contributes to understanding the causes of cyber abuse or soft security, which refers to an online behaviour that aims to upset or harm other users in many ways for a variety of reasons which is listed and reviewed in the following Chapter 2. The key contribution of the thesis can be summarised as follows:

1.4.1 Understanding perceptions on content moderation

Content moderation has become a de facto necessity on platforms that enable user contributed content in order to conform with standards of acceptable use required by regulators and/or social norms. This, however, has encountered multiple challenges in

the field of social computing. Specifically, creating better guidelines for content moderation to cope abusive behaviour across online communities. The analysis of online social behaviour can articulate communication gaps by exploring causes of dynamics between moderators and contributors to design superior social system.

Prior work has characterised moderation system in terms of process and policies for intervention using mostly qualitative research analysis. This research is conducted to identify the key perceptions and motivation of content moderation using mixed approach empirical analysis to study the differences between users' perceptions and daily actions or activities.

Due to the scale of content on such social platforms, many providers have adopted a mixture of automated and human moderation processes. A considerable amount of research has been undertaken to understand how moderation influences the behaviour of end users. However, the interplay between social understandings of norms, the development of platform policies and their enforcement on contributors are less well understood. A considerable amount of research has been undertaken to understand how moderation influences the behaviour of end users.

To begin to explore this perspective in Chapter 3, a survey was undertaken of moderators on the popular discussion site Reddit. The analysis revealed that moderators on Reddit view themselves as members of their community and work hard to both guard against violations, but also with contributors to enhance the quality of their content. Also, suggested that moderators work at different times to contributors and reflect upon the implications of these findings for the design of moderation systems.

1.4.2 Online vs. in-person behaviour in conversations

Chapter 4 investigates the potential for online settings to disinhibit abusive behaviour that involves promoting individuals to behave potentially differently between online and in-person discussion or public and private discussion as a group. Understanding the similarities and differences between online and in-person settings will lead to filling the gap between conversations in different settings which contributes to the computational social science field. In particular, proposing effective methods and analyses to uncover the implications related to detecting abusive content. This can be achieved by

finding appropriate textual features to detect abusive content and investigating the key factors of behavioural changes between online and in-person group discussion through both qualitative and quantitative characteristics of analysis.

Extensive work was carried out to apply social theories in content analysis. This work, however, investigates the correlation of behaviours that can lead to abusive content between two distinct types of communications namely in-person and online.

The chapter investigates the impact of online and in-person settings on abusive factors in small group discussion within a Software Engineering Team Project course. Using qualitative and quantitative methods, (N = 67), the study examines how they interact and behave with one another during an academic year. The analysis suggests that the online setting encourages behaviour that can eventually lead to abuse within discussions. In particular, the online discussion setting fosters participant behaviours (less hedging, more extreme sentiment, greater willingness to express personal opinion and straying from topic) that are known to increase the likelihood of abusive behaviour. The proposed classification model is able to accurately detect online (public) vs in-person (private) conversations based on abuse and politeness, and sentiment features. A measure of peer evaluation of conversation was used to understand stimulated behaviour among groups in terms of abuse related context between in-person and online conversation. Furthermore, the findings show how consensus building factors can influence discussions in different settings. The chapter concludes by discussing the theoretical implications.

1.4.3 Disagreement and abuse in online discussions

Disagreement often involves multiple forms and stages of argument. These arguments in an online discussion can often lead to abusive content. In particular, the disagreement scale that contributes to both computational social science and social computing can offer another mechanism for social platform designers to implement assistive methods to distinguish between polite and abusive disagreement. Also, the statistical analysis for a social network platform reveals causes of insults or severe-behaviour in different online communities.

Prior work has attempted to adapt approaches of abusive text detection from the fre-

quency of terms in the context then used for auto moderation. However, not all cases of abusiveness are captured appropriately if we do not know how to differentiate between disagreement and abuse in a given context between two or more users.

In Chapter 5, more than 200k of comments were examined from discussion threads from the top five subreddits on Reddit. The chapter showed how disagreement and abuse interrelate. Using the set of definitions of disagreement levels, then different types of messages were classified accordingly and what features correlate with them to build a disagreement classifier that aims to uncover the ambiguity between the abusiveness and disagreement in discussions from 5k tagged comments. This process allowed patterns of abuse and patterns of disagreement to be compared. The findings showed how disagreement can be detected in the context.

Nevertheless, the findings of the above contributions imply some suggestions and implications that addressed the research gaps in social computing and computational social science research domains. In particular, understand the key aspects of content moderation that create social obstacles and impact the behaviour of contributors in social platforms.

1.5 Thesis Overview

The thesis analyses the way humans interact with social platforms from disagreement and their point of view. Therefore, the chapters can be summarised as follows:

Chapter 2 reviews related work to this thesis. Mainly, the research body from the literature that investigated similar claims using a variety of methods and approaches including machine learning, empirical analysis, case studies and qualitative research.

Chapter 3 explores moderator behaviour on intervention strategies and motivations. The chapter examines both contributor and moderator activities at the macro-scale to uncover causes of conflicts and perceptions of moderators.

Chapter 4 explores the aspects of abusive behaviour online and in-person for small peer-groups discussions. The chapter includes empirical analysis at the micro-scale

from textual conversations and both pre and post surveys.

Chapter 5 examines large-scale data from Reddit to investigate factors of disagreement and abuse, and develops a measurement of disagreement scale. The scale is used to distinguish between polite and abusive disagreement using five defined disagreement levels. The chapter discusses possible factors that can lead to abusive context.

Chapter 6 discusses the interplay between abuse, disagreement and moderation that is conducted in main chapters, and compares it with the literature. In particular, considering the anticipatory approach to moderation-based online communities to reduce abusiveness.

Chapter 7 concludes the research objectives and answers the questions introduced in the introduction chapter and suggests potential research directions guided by this thesis.

1.6 Publications

The following list of research work were published during this dissertation:

1. **"Understanding Abusive Behaviour Between Online and Offline Group Discussions"**. Abdulwhab Alkharashi and Tim Storer and Joemon Jose and Andrew Hoskins and Catherine Happer. *In proceedings of the 37th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI'19)*. May 2019.
2. **"Privacy in Crowdsourcing: a Systematic Review"**. Alkharashi, Abdulwhab and Renaud, Karen. *In proceedings of the 21st Information Security Conference (ISC'18)*. September 2018.
3. **"Vandalism on Collaborative Web Communities: An Exploration of Editorial Behaviour in Wikipedia"**. Alkharashi, Abdulwhab and Jose, Joemon. *In proceedings of the 5th Spanish Conference on Information Retrieval (CERI '18)*. May 2018.

Chapter 2

Literature review

2.1 Introduction

This chapter, a review of the state-of-art relevant research, abusive behaviour, on-line social settings. In particular, the characterisations that prompt abusive content. These characterisations of abusiveness are often crucial elements to the fields of computational social science and social computing. The computational social science field seeks to understand social phenomena through several computational methods including empirical analysis of large-scale and structured data. The social computing field is an intersection between social science and computer science that is often used to describe design or challenges of social systems. In particular, investigating the key issues of moderation systems in online communities that are more likely to promote anti-social behaviour in order to design better intervention systems for social platforms.

The chapter is structured as follows. Firstly, Section 2.2 describes the primary definitions related to forms of online abuse on social platforms. This provides a conceptual framework for the later sections, and multiple parameters of abusive behaviour, the different definitions given in the literature for the different forms of online abuse. Section 2.4 reports the causes of online abuse. This is the factors that prompt users to commit abusive behaviours. Sections 2.5 reports the methods of abusive content in online social platforms and investigate about such behaviour. The following Section

2.3 surveys the content moderation actions. Specifically, the perceptions and expectations of the moderators and contributors in online communities. This will facilitate to bridge the gap between intervention designs and human-decision on content moderation. The chapter concludes in Section 2.6 by summarising the major research gaps discovered in the literature.

2.2 Definitions and forms of abusive behaviour

There are several definitions and forms of abusive behaviour across online communities. There are multiple forms of abusive online, e.g., trolling, sockpuppet, cyberbullying and cyberstalking, swatting, non-consensual (sexual assault), and doxing [41]. In some circumstances, any forms of verbal abuse online could lead to a physical harm or violence [42]. Summary of definitions from the literature for these forms of online abuse is presented in table 2.1. The definitions for variety forms of online abusive behaviour are essential to review in order to develop a broad perspective of the online communications issues in the thesis.

Trolling. The term trolling has been used since the early 1990s on discussion boards such as Usenet to describe a form of online abusive behaviour that is disrupting an online discussion between individuals by posting abusive or off-topic comments [43]. Nevertheless, not all trolls aim to harm communities or people. For instance, trolls may ask naive questions or post over-discussed topics to deceive newcomers for enjoyment purposes or as they are called ‘stuck in the LOLs’.

Coles and West [8] defined trolling as taking pleasure of intentionally upsetting others online when posting either off-topic or inflammatory comments on social platforms. The authors examined the collected comments that related to trolling terminology, and identified four aspects or characteristics of trolling behaviour : (1) recognisable (2) reminiscence (3) vigilante (4) inconsiderate. These terms suggest that trolls are mostly identifiable and have tendency to do harmless actions. The term vigilante is used to describe the effects of trolls and counter-normative behaviours.

Also, trolls hold anti-social characteristics. Golf-Papez and Veer [44] adds further characteristics stating that trolls seek for entertainment and intend to avoid frequently tar-

Author(s)	Scope	Definition
[8, 44, 45]	Trolling	Post destructive comments online to upset others.
[47]	Sockpuppet	Form of trolling that recruit troll-bots to commit trolling activity.
[48, 49]	Cyberbullying	Desire for making fun of people online.
[50]	Cyberstalking	Track victims online to terror them.
[51]	Swatting	Making fake calls to law enforcement to report false event of victim's address.
[52, 53]	Doxing	Distribute sensitive information of victim for harassment and threats purposes.
[54, 55]	Non-consensual	Blackmailing the victim to threat for publishing intimate relationship online.

Table 2.1: Summary of definitions for most common forms of online abusive behaviour highlighted from the literature.

getting one user to upset or conduct any negative behaviour online. Donath et al. [45] states that trolls enjoy what they do while producing misleading information on social platforms. Hardaker [46] claimed that trolls mostly rely on offensive language and hate speech to the pace of a discussion. Unlike vandals, trolls make a clear intention to disrupt particular community or user, whereas vandals do not necessary target specific online group and may have other purpose [19].

Another form of trolling appears to target online gaming communities; these are called griefers. These characters aim to upset online multi players or pose financial harm to gaming industries and break the communities' rules (hate speech and offensive language) [56]. Achternbosch et al. [57] studied the rationales of online griefing from both the attacker's and victim's perspectives. The study utilised an online survey and reported that victims of griefing who score a higher ranking in the game are less likely to encounter problems in real-life. Regardless of the causes of online abusive behaviour, victims of online abuse cannot easily differentiate between harmful and harmless threats.

Sockpuppeting. Kumar et al. [47] defined sockpuppeting as an automated user accounts

designed to post comment or pages, reply to comments and vote. This approach is also considered a form of trolling behaviour. The authors examined nine discussion communities to study the sockpuppet's posting behaviour and how it can be identified. The study suggest that sockpuppet lean to be less active in terms of posting discussions per thread, yet post higher number of replies to other comments. Sockpuppet also use less words in posts or replies in general, mostly contain personal pronouns, and receive significant amount of down-votes by the community members.

Cyberbullying. Watts et al. [48] defined cyberbullying as persistent behaviour that aims to harm or embarrass an individual via social platforms for fun reasons. The authors report that the main causes for the increase of cyberbullying is the abundance of the internet accessibility in phones or other portable and electronic devices. Wolak et al. [49] investigated the characteristics of online harassment between peer-group and online young adults of age 10-17 years. The study utilises a telephone survey of 1500 participants and concludes that cyberbullying is an extension of bullying that takes place between people who know each other physically. Patchin and Hinduja [58] examined the correlation between self-esteem and cyberbullying of young students who completed an online questionnaire. The study concludes that people with low self-esteem are more likely to encounter cyberbullying.

Cyberstalking. Cyberstalking is another form of cyberbullying in terms of intentions of harms except that cyberstalking includes tracking the victims to send threat messages to make them feel frightened [59]. Spitzberg and Hoobler [50] conducted three experimental studies to understand the motive and measurement of cyberstalking behaviour. At least one third of the samples of college students' responses reported that they experienced cyberstalking. The research's findings suggests that cyberstalking is often linked to interpersonal terrorism.

Swatting. Swatting is defined as a form of abusive behaviour that seeks to make particle jokes on the victim by spoofing caller ID to make phone calls to the special force of policing to report a false incident, (e.g, bomb or weapon shooting) and provide the address of the victim for revenge or humour reasons [51]. Swatting dose not occur online directly, yet as result of online abusive behaviour, it appears in in-person situations. Benderev [42] reports that a young male committed a crime of swatting by reporting a false shooting incident over the phone and linked it to the location of victim's home,

which caused a loss of the victim's life. The author also reports that officials are unable to tackle this issue of fake incident reporting or swatting due to the high volume of daily calls received from multiple citizens. Wu [60] argued that some online communities that lack active moderating activity or non-monitored community (e.g, Reddit, 4chan and 8chan), promote abusive behaviour including swatting, specially against gender and race.

Doxing. Dropping docs known as doxing involves in stealing user's identity or sensitive information to publish over the social networks sites as a result of conflict or revenge to the victim [52]. Snyder et al. [53] investigated high volume of text files posted in four social platforms to develop a tool that can capture doxing activities. The author report that doxing is linked to equity and revenge reasons mostly. The analysis suggests that developing a mechanism that inform users when their accounts is compromised by an attacker is essential.

Non-consensual. Citron and Franks [54] defined non-consensual or revenge porn as a continuous threat of media files that show intimate relationship of the victim and used to either break relationships or keep them silent. Kamal and Newman [55] characterised the non-consensual as cyberharassment behaviour and imply that it can lead to mental health issues. Suzor et al. [61] argued that constructing effective policies on web platforms that offer access to abusive materials is a challenging task.

Pietrangelo [62] explains that verbal abuse is a repeated action that tend to humiliate and petrify individuals. Both verbal abuse and emotional can lead to a physical abuse eventually. Pietrangelo provides several examples where an argument can contain an is acceptable disagreement or verbal abuse. For example, disagreement may imply that a polemicist's objective in an argument is to give the second person; who may or not have an opposite opinion, opportunity to express the discussion statement even if s/he is upset. Also, the argument should not contain any name-calling, offensive language, or direct insult to the characteristics of a person rather than the actual argument. There are several signs and early stages of abusive behaviour in interactions (e.g, name-calling, condescension, criticism, degradation, manipulation, accusations, circular arguments, gaslighting and blame) [62]. Steele [63] investigated the correlation between name-calling and compliance among housewives on two experiments and found that name-calling which uses negative judgement or name can lead to com-

pliance behaviour.

Blau [64] proposed a framework of Social Exchange Theory associated with marketplace and human interaction to understand the social behaviour of relationships between individuals. The theory uses economic terms such as cost and reward in the context of human behaviours. The cost term implies that a person is receiving negative social values including time or money and reward is positive social value such as support as so on. The theory suggest that people are more likely willing to maximise their reward profit in relationships by reducing the elements of punishments to receive positive relationship. Cohen and Felson [65] presented the Routine Activity Theory that aims to uncover the interplay between high crime rates and social activity. In particular, the theory suggests that households with working occupants are easy target for burglary whenever they are at work. Also, homes with no secured methods to protect the property are more vulnerable. The theory has been implemented into multiple disciplines including information security [66] and online negative behaviour [67]. Gainsbury et al. investigated the factors of online negative behaviour. The study utilises an online survey of Australian internet users and concludes that online abusive behaviour is unpredictable, yet it can be correlated with a particular event. Sofield and Salmond [68] reported that most intentions to leave cases of nurses in large hospitals were related to verbal abuse.

2.3 Moderators and their view point

This section reviews the related work concerning the practice of moderation and its impact on contributed content. This existing research literature helped shape the selection of exploratory questions that contribute to moderation systems and social computing.

2.3.1 Moderation impact on contributors

Several researchers have investigated the impact of moderation activity on contributors. For example, Chandrasekharan et al. [69] found that Reddit's decision to close a

number of subreddits for violations of its acceptable use policies resulted in the participating users substantially improving their behaviours. Users largely stayed on the site, but did not transfer their unacceptable behaviour to other subreddits. This suggests that behaviour is significantly influenced by others within an online community and that users adjust their behaviour to be acceptable within multiple communities.

Chen et al. [70] developed a theoretical model of contributor behaviour in response to moderation and contribution rating. Chen et al. [70] argued that the model shows moderation is generally beneficial in improving the quality of content in a discussion. In addition, they suggested that strategic contributors will operate strategically, providing high quality contributions to boost their reputation, before exploiting their reputation. Another study [71] reports that effects of peer-moderation are useful in enhancing the conversations among participants.

In later work, Chandrasekharan et al. [36] studied the dissemination of value norms across Reddit. They demonstrated that certain norms are universal, such as insulting or abusive behaviour. Other norms are widespread, but not universal, such as avoidance of criticism of moderators. Finally, micro-norms, such as the use of particular language styles are highly specific to particular subreddits. The authors contend that the discovery of common norms (at the macro and meso level) has implications for the design of automated moderation systems, since this reflects commonality between different communities. Cheng et al. [72] explored how community feedback influenced quality of contribution on online discussion platforms. In contrast to the work of [70], they found that negative feedback had a significantly negative influence on the future contributions of the same authors. In addition, the negative feedback also influenced contributors to rate other content more poorly.

Wright [73] studied the impact of authority on moderated behaviour, revealing that online discussions moderated by administration officials appeared to be more civil behaviour by participants than is observed in unconscious, unofficial online discussion.

Community rules can interplay the degree of casual norms rather and create assistive methods. For example, Butler et al. [74] investigated the complexity of rules and policies in Wikipedia and reported that platforms that enhance affordances and employ side-walk strategy are capable of offering wide range of supportive designs and activities.

Earlier work [75] on governance in social networks proposes the significance of ac-

cepted practices in directing behaviour, yet additionally realise that the trouble for newbies adapting norms can prompt relatively high cases of leaving an online community and increased the retention rate. The authors concluded that communities should consider attempting to intervene ahead to make acceptable interaction between newcomers and existing users. Similarly, Choi et al. [76] examined the socialisation effects for conventional groups in Wiki participatory projects. The analyses revealed that newcomers are most likely to contribute less, yet active users become less active over time. The study concluded that social tactics interplay the the roles of community users online.

Social norms on online sub-communities are complex due to the vast types of communities with multiple cultures and subjects which may show down moderators' decision process. A few norms, nevertheless, are embraced from the overall social setting. For instance, derisive labels demonstrate discourteousness. On the other hand, a few behaviours are shared over the web, similar to either users are being abusive or seeking fairness across multiple communities [77].

The unpredictability of content moderation and the obscurity with which online communities handle make it hard to look at how moderation systems handle complex use-case scenarios. However, throughout the past decade, multiple attempts were made in research into understanding the various parts of content moderation, normally utilising hypothetical or subjective methodologies. For example, Grimmelmann [78] argued that moderation systems can be executed in a way that comply social settings and expectations of community norms between control and freedom of moderation actions. Similarly, prior work [79] conducted an online survey with content moderators to understand the experiences from their points of view of multiple levels beyond freedom of positing controversial topics. Also, [36] examined about three million removed comments by moderators on Reddit to investigate the impact of social norms at different levels. Both researches [36, 79] suggested a modified moderation system that supports community norms.

Crawford and Gillespie [80] investigated flagging mechanism and its motives from the user perspective for content that is hostile or that abuses the community rules. The authors argued that flagging can promote bullying behaviour. In particular, users may flag comments of their fellow peers for enjoyment or prank reasons. Also, flagging can

leave users wondered about the cases for comment removal.

Prior work explored circulated content control, which includes depending on total of users appraisals to assess a comment or post [81, 82]. Numerous researchers [83, 84] have investigated how Wikipedia contributors by using Talk pages designed for discussing update or changes on a particular topic and concluded that the roles of online productive community can interplay the level of contribution among users.

Jhaver et al. [85] argued that new systems and approaches are expected to unravel the human from the platform. Without such techniques, this should depend on evaluations of which human involvements are likely to emerge [86]. There are legitimate functional explanations behind social applications to settle on such plan choices, however, much of the time these either dark or lose significant parts of the fundamental human behaviour. Evaluating and, if conceivable, amending for these capacity and access approaches ought to be a piece of the dataset detailing and initialisation process.

Notwithstanding numerous researches in content moderation that have been studied to understand the challenges related to content moderation, there are huge of spammers and internet bots taking on the appearance of typical people on all major online social networks. Binns et al. [87] investigated a challenge on content moderation relates to the gender bias and imbalance and found that females tagged more toxic comments than males. Jiang et al. [88] conducted an interview with 25 active moderators on Discord platform to describe another challenge about voice-based moderation. The authors suggested that this approach can lead to false accusations and disruptive noise from users. Seering et al. [89] presented three procedures that contributed to moderation engagement and concluded that moderation systems need to adopt user-driven approaches that is caused by influential aspects.

Besides, numerous conspicuous people keep up internet based life accounts that are expertly figured out how to make a developed picture or even carry on in order to deliberately impact different users. It is as of now difficult to precisely evacuate or address for by far most of such contortions.

2.3.2 Behaviours and perceptions correlated with moderation

Research has explored how social norms [36] and hate speech [69] can affect the banning activity on Reddit. Moderators on Reddit have a variety of privileges for regulating the subreddits they moderate ¹. The list of actions that they have are including removal, approval, (mute or suspend), add flair to the title, tag posts NSFW for inappropriate content, and pin posts. Conventional methods of content moderation have relied mostly on automated text detection by extracting lexical features to build supervised and unsupervised classification models for capturing an appropriate content. Nevertheless, these approaches can not guarantee combating abusive content and behaviour. In particular, moderators and contributors behave differently across distinct online communities, which may factor the conflicts between contributors and moderators. Therefore, knowledge, the relationship between review and intervention on moderation is still under-investigated.

Removing a comment or post can be clear case when the user is violating community rules, e.g, nationalism or racism [90]. Yet, there are few case scenarios where the comments or posts will be removed from a community due to disagreement between users and moderators led by spiral of silence factors [91, 92]. This is a critical issue since while users are making en efforts to belong to a community. Some discussion platforms, e.g, reddit has a removal reason tool [93] that permits moderators to either select pre-defined reasons or add a new reason for removing a post. However, it may become an ambiguous decision to adapt without understanding the correlation between reasons for intervening and topic genera.

The expansion and implementation of moderation rules must be transparent to sustain censorship on daily basis or it may lead to perception of bias among individuals [94]. A vast issue of trust has to be placed in the opinion of the moderator not to improperly review posts and manipulating affordances cause lack of understating the rules of community [95]. Researches of online public trust [73] have revealed that online discussions moderated by administration officials appeared in more civil behaviour by participants than is observed in unconscious, unofficial online discussion. Another study [71] reports that effects of peer-moderation are useful in enhancing the

¹<https://www.reddit.com/wiki/moderation>

conversations among participants. Nevertheless, the question may arise is who shall moderate discussions and why? Is it be independent moderator, an employee or agent of an organisation, a member, or 'impartial' auto-moderator? Since it is challenging to comprehend when posts are being reviewed, these kinds of problems prove complicated to answer.

Conceptualising around community and how norms develop can affect the decision of intervention. For example, Lee and Lee [96] studied how media use and demographics factors interplay the use of online communities between users and non-users. The study recruited 327 survey participants (41% online community users). The study's findings suggested that online social networks may affect the correspondence capacity of how people interact more than face-to-face communication.

Gangadharan [97] looked at broadband selection programs at network-based and open foundations in the US to investigate issues related to privacy and surveillance among internet users in educational environments. The study reported the analysis from both groups and individuals' discussions and in-class observations between forty student and five instructors. The findings suggested that students in introductory level are more likely to encounter poor privacy enhancement. Cantijoch et al. [98] conducted a study that aims to reveal motivation of civic action sites that promote empowerment in local communities. The study relied on a web source for the published survey data namely mySociety (N = 6239). The findings suggested that inclusion in aggregate as opposed to singular methodologies can help reducing problems to increase people engagement. Sumner et al. [99] leveraged a functional approach to identify the reasons for the use of Like button in Facebook recruiting 156 users. The findings claim that frequency of Likes and the users' interpretations can interplay the reasons for using Like buttons.

Ragnedda et al. [100] proposed a digital capital index that aims to measure multiple contexts from an online survey of 868 UK citizens. The results of the proposed model suggest that socio-economic and socio-demographic patterns including age, educational level and income are likely to affect the digital capital index. Schrader et al. [101] presented a conceptual study using mixed methods to understand the key factors of users' actions on the game League of Legends. The study uses Reddit to post general queries about the familiarity about the game and collect the responses, followed

by development of the scale to use on the survey to collect data from multiple social network resources. The results presented suggestions for instrument improvement in constantly developing settings.

Cenite et al. [102] examined the differences of ethical beliefs and practices on personal and non-personal groups of bloggers. The study uses an online survey of (N = 1224; non-personal bloggers 27%) recruited participants after conducting a focus group meeting of 70 bloggers to investigate ethical concerns that bloggers are most likely to encounter. The findings suggested that the two groups of bloggers were distinct of what they present in their blogs—the two samples reported that attribution is generally significant and responsibility least significant.

Poor [103] explored the motivations and sense of online gaming moderators by using both survey data and interviews analysis of 111 respondents. The findings implied that young and old moderators contribute more than one moderating game and have strong sense of community motivation. The moderators have expressed that they most of the time collaborate with one another and have an ambition to join the gaming industry.

2.3.3 Research gap

The conflict between contributors and moderators have been presented from different aspects including both theoretical and practical models. Nevertheless, the expectations of intervention activity and motivation in abusive content is still speculative. Also, the role of moderators and how they perceive themselves has not been studied clearly. In particular, we do not know if the activities of contributors and reasons of interventions from moderators create more conflicts or not. Also, it is unclear when moderators and contributors become more or less active during the week, and why moderators react differently and take action towards particular comment.

2.4 Causes of abusive behaviour online

Abusive behaviour on online communities can be caused by numerous reasons. Understanding the causes of behaving against community norms and expectations is crucial

to building proper method to prevent or reduce such behaviour. This section reports studies related to the cause of online abuse on social platforms including: community reactions, situational factors, anonymity, online vs in-person and social norms.

2.4.1 Community reactions

Community reactions (i.e., votes and replies) play a significant role in altering user behaviour while contributing. Cheng et al. [31] examined a large amount of posts and votes of four news communities to understand the feedback (i.e, down-vote and up-vote) on posts that impact users' contributions and behaviours. The authors found that users tend to post more comments when receiving negative votes by their fellow users and post lower quality of content after negative evaluation. Positive votes, on the hand, did not impact the vote and post behaviour quality. Users who did not receive any votes are more likely to leave the community. Similarly, Tan and Lee [104] investigated the aspects of interaction across multiple online communities on Reddit and computer science bibliography website (DBLP). The DBLP was used to find published papers in computer science conferences for multiple research areas. The experiment targeted three primary objectives of user activity including community reactions and language aspects. The study targeted active users who posted more than 50 posts and authors who published at least 50 papers. The findings suggested that while users are attempting to adapt community norms including the language and diversity of topics, positive feedback of up-votes caused existing (loyal) users to contribute more by posting in smaller communities. Stroud et al. [105] studied the relationship between reaction buttons; like, respect and recommending on commenting section political news articles. The authors reported that using respect button is more appropriate for eliminating social basis on political debates as a result of agreement. Kumar et al. [106] investigated more than 36K subreddits on Reddit to learn about the the conflicts. The analysis suggests that active users are more likely to start with negative comments, yet less active users spread their behaviour across wider number of subreddits due to the influence of social interaction. Negative behaviour led to extreme conflict and direct communication between attackers and victims. Similarly, Crandall et al. [107] proposed a model to understand the relationship between social selection and social influence In Wikipedia , selection behaviour is when people seek other people who

share similar interests and values. Social influence refers to people, share similar social status, religious beliefs, interests and so on. Using cosine similarity measure to find the similarity between users' interaction, Crandall et al. reported that community feedback play a significant role in affecting the users' contribution prior and after social influence and selection. Also, similarity in behaviour promotes social interaction amongst users. These results are caused by socio-economic factors.

2.4.2 Situational factors

People may behave differently according to particular situation or circumstances which may impact the way they interact in social platforms. For example, Cheng et al. [108] conducted an online controlled-experiment to see whether mood has a direct effect on posting abusive comments on an online discussion community. In particular, designed an online simulated discussion to examine users' comments on news communities after completing a qualification test. The finding showed that users are more likely to receive negative comments and trolling behaviour when they have received low scores in the qualification test. Thus, they concluded that situational factors may play a significant role in shaping users' behaviour online. Cheng et al. [27] investigated the factors of anti-social behaviour online on the commenting section for three active news communities. The study examined characterise users in two groups: never banned users and future banned users, the study suggested that banned users are more likely to post unrelated content and cross-posting comments or posts that belong to other users. In addition, when users are banned, they receive aggressive feedback by the community members. Buckels et al. [18] claim that trolling behaviour is due to entertainment reasons (known as grievers in online gaming communities) and have strong ties with some personality traits. The study drives the analysis from online surveys collected for two replicated studies. The authors suggest that further investigation is needed to clearly understand the correlation between trolling engagement and behaviour within rigorous empirical analysis.

2.4.3 Anonymity

Anonymous trolling affects multiple online communities [109]. Tucker [110] investigated the causes of abusive comments in an anonymous evaluation survey. The study gathered more than 17K responses about the overall feedback of teaching staff and course structure of which all have more than 30K comments from students. The author reported that there were some abusive comments that addressed directly to teachers and learning experience. The author concluded that submitting comments anonymous online can, prompt students to provide abusive content, and suggesting that institutions must adopt a proper strategies to combat the abusive behaviour, e.g, educate students and teachers about sharing feedbacks effectively. Black et al. [111] studied the influences and concerns of anonymous social platform namely Yik Yak on US colleges. The study reports that vast majority of online posts contain offensive language, particularly, profanity vulgarity and sexual assault. The study concludes that social platforms that support anonymous posting, and promote negative behaviour are more likely to encounter forms abusive content. Suler [16] presented six key factors that contributed to online disinhibition, that users are willing to interact online differently than in-person. Lapidot-Lefler and Barak [11] conducted an experiment to examine online disinhibition, and found that lack of eye-contact can lead to disinhibition.

Anonymity on discussion threads gives a platform to users to discuss freely about variety of topics with one another on an equivalent balance that is both likely and expected to occur. Anonymous users may often contribute less online, yet provide high-quality content on productive platforms such as Wikipedia For example,[112]. Kang et al. [113] carried out a user-study to comprehend the motives of using social platforms that support anonymous identity. The authors report that participants feel that anonymous communication allow them to express thoughts freely, openly and safely.

2.4.4 Online vs in-person

Prior research argued that online interaction might not be the same as in-person for multiple factors. Theses factors imply that online conversation may foster the disinhibition, which encourages individuals to carry on behaviours online– that is not

necessarily would do in face-to-face interaction [16]. Also, some online social platforms use asynchronous form of communication which allows individuals to take time and think about the responses online instead of responding in-person communication. McCully et al. [114] investigated the differences between online and in-person interactions through interviews and content analysis of small number of users who post on Everything2.com platform. The findings suggested that in-person interaction encourages strong relationships and disinhibit participation of contribution on online content.

Williams et al. [115] built on previous study designs using a ball tossing game to study the effects of ostracism on different mixes of participants (online, in-person). In particular, to investigate factors of exclusion in online communication. The study utilised controversial topics for conversations. Moreover, participants were asked to complete an assessment to evaluate negative effects. The analysis suggested that disagreement is led by control and self-esteem factors. The authors concluded that people who experience social exclusion are more likely to encounter negative emotion and less impact of ostracism. Williams et al. [115] were the first to study online and in-person, and have given some significant data in regards to the acceptance following rejection of interactions. Filipkowski and Smyth [116] conducted further two experimental studies to examine the discrimination of ostracism among online and in-person conversations. The authors reported that online communication may permit individuals the ability to effectively adapt to the fact of being rejected. Both online and in-person expectation of ostracism can impact people in identical way. The studies concluded that online communication prompt self-esteem and in-person communication discourage self-esteem over time.

Cleary and Walter investigated the comparison between email and in-person interview meetings teenage with mental disorders and concluded that face-to-face interaction is far effective [117]. Further analyses investigated the equality of meeting information gathered in-person versus online. The findings show that young people favour online techniques. Mason and Ide [118] investigated the differences between face-to-face and email interviews for youthful people. The participants reported that they were passionate about the email interviews rather than conventional face-to-face interviews, yet felt that a synchronous email interviews is way slower than synchronous commu-

nication such as text messaging or in-person interview.

Research about group discussion impacts the manner in which users respond and think. For instance, an exploratory investigation [27] analysed the impact of viewing material concerning child abuse. The oppressive behaviour in social networks can be as in-person behaviour which is framed by hostile acts and digital harassment [119, 120]. Huang et al. [121] report an interview analysis to see how people behave in event invitations online and offline. The authors report that the setting of the event and strength of relationship between invitees play an essential role in shaping people behaviour in event invitations.

In light of the above observations, in-person communication lean towards confirmation of social perception, so individuals will, in general, adjust to online social settings that may interplay social interactions.

2.4.5 Social norms

Social norms are regulations and principles that are comprehended by individuals from a gathering, and that control and additionally compel social behaviour without the power of laws [122]. Notwithstanding normally acknowledged principles of behaviour, standards incorporate guidelines restricting unsatisfactory social practices, for example, restrictions against inbreeding or child murder, and laws or norms for lead built up by a legislature or chose body [123]. Norms, however, play significant role in shaping our behaviour identified with increasingly daily exercises also, from how uproariously one ought to talk on a mobile phone in an open space, to what the fitting dress is in various social circumstances.

Several authors have investigated how users' behaviour is influenced by social media. Heider [124] presented the Balance Theory that suggests that attitudes among individuals are based on a balance sentiment relationship, i.e., dislike or like of something. Some authors have adopted Heider's Balance Theory to examine the impact of social media on behaviour. For example, Nakanishi et al. [125] presented an experimental study to examine group effects leveraging Balance Theory. The authors reported that agreement promotes positive sentiments and negative sentiments for disagreement. Posegga and Jungherr [126] examined political event hashtags to characterise political

debates on Twitter. The findings showed that political debates had strong correlation with public agenda which is caused by similarity of interests and objectives.

In the context of Internet privacy concerns, Tufekci [127] investigated the differences between public and private disclosure in social networking sites. The study reported that users overcame their privacy concerns by modifying their profilers using nicknames. Evaluating and, if conceivable, amending for these capacity and access approaches ought to be a piece of the dataset detailing and initialisation process.

Uski and Lampinen [128] argued that users in social network sites encounter social attitudes or norms as a result of self-impression and acceptability of cultural background or beliefs. Salmivalli and Voeten [129] investigated the association between attitude and group norms of young students showing abusive behaviour cases. The analyses suggest that group norms are directly affected by young females and attitudes by young males. These norms develop out of association with others; they might possibly be expressed unequivocally, and any assents for veering off from them originate from individual individuals from the social gathering, not the lawful framework. Cialdini et al. [130] argued that norms fluctuate to the degree in which behaviour is approved or disapproved due to influence perceptions. Ajzen et al. [131] proposed the Theory of Planned Behaviour model to predict human behaviours and intentions to involve in particular actions. The theory addresses five aspects that justify human intention and behaviour: attitudes, subjective norms, behavioural dominance, behavioural intention and social norms. Cialdini and Goldstein [132] states that group norms are more likely to influence behaviour in compliance and conformity situations. In particular, people may mimic behaviours surrounded by people close to them if they have no intention of an action. On the other hand Goldstein et al. [133] investigated descriptive norms in relatively paired individuals and found that normative action is crucial to determine group behaviour that is led by situational factors.

2.4.6 Research gap

In summary, this section provides an overview of guidelines of behaviour in online discussions in relation to community reactions, situational factors, anonymity behaviour, disinhibition, and social norms. There is a significant assemblage of work of online

communications covered in this section. Yet, little is known about what causes abusive behaviour online in relation to social norms. In particular, we do not know if the language used online is different from in-person conversation. On the basis of the literature, we hypothesise that the nature of the language used online is qualitatively different to that used in person and thus more readily leads to the introduction of abusive content and behaviour. We anticipate that identifying different forms of disagreements in these contexts will reveal different patterns of behaviour in in-person and online settings.

2.5 Methods for tracking abusive content, performance of automated tools

There are several automated approaches to capture abusive behaviours across online communities. The approaches are statistical and rely on pre-existing activity of users and other approaches rely on contextual aspects or community reaction. The approaches also apply machine learning techniques to facilitate and expedite moderation process that requires human decision-making to prevent violations of community rules. This section demonstrates prior work on three primary dimensions for identifying abusive behaviours and compare its performance.

2.5.1 Contextual-based detection

Most previous work on identifying abusive content concentrated on natural language processing (NLP) techniques along with text classifications which may also require human annotation tasks to build datasets in order to serve the purpose of building better classifiers. In particular, prior researches developed several approaches for abusive content that is frequently identified with sentiment and opinion [134]. A summary of the selected proposed method of detecting abusive text is shown in Table 2.2.

In most cases, data scientists need a pre-labelled or classified input class of the raw data to train a classifier or build a text predictive model. One way to achieve this process is to post list of data annotation tasks on crowd-sourcing platforms to recruit workers

Author(s)	Feature	Model	Dataset	Accuracy (%)
Khan et al. [136]	Offensive	TOM	Twitter	85.7%
Sood et al. [137]	Offensive	SVM	News	92%
Davidson et al. [25]	Hate Speech	One vs Rest	Twitter	61%
Warner and Hirschberg [138]	Hate Speech	SVM	Yahoo	94%
Founta et al. [139]	Sarcasm	Interleaved	Twitter	97%
Dinakar et al. [140]	Sexuality	JRip	YouTube	80.2%

Table 2.2: Summary of top performance for the proposed automated text detection methods on content abuse from the literature.

to complete particular task. The process of building the predictive model is a type of supervised learning that uses known machine learning algorithms, (e.g., Support Vector Machines SVM, Random Forest, Naive Bayes, Logistic Regression). Typically, these algorithms are useful in applications such as spam filtering, crime prevention, social media analysis and abusive content detection, Nobata et al. [135] developed a detection approach to identify abusive language that embodies hate speech and profanity on Yahoo news and Finance data. The corpus contains labelled data by crowd workers in Amazon Mechanical Turk. The approach consists of training the classifiers Vowpal Wabbit machine learning tool for binary classification tasks. The performance of prediction achieved 82% at Yahoo news datasets in detecting whether the input body contains abusive or clean text.

Khan et al. [136] proposed a method for mining the content from six Twitter datasets by presenting a sequence of steps on pre-processing phrase to preform text classification tasks. The steps aim to clean text (tweet) prior building the classification model by detecting and analysing slang or abbreviation to ameliorate text classification task. The authors reported that this approach achieved on average 86% accuracy rate across the six datasets.

Sood et al. [137] proposed a profanity recognition framework utilising words from the subject as a dishonest lexicon. They additionally applied edit distance metric to identify slang text and Support Vector Machine (SVM) to detect the profanity text. They used a set comments from a social news sites to assign annotation tasks by crowd workers on Amazon Mechanical Turk to decide the existence of profanity in a given comment. The overall performance of the classification model reached 92% accuracy score.

Another critical type of abusive content is hate speech. The main difference between hate speech and offensive language is that hate speech is, intended, to offend a particular racial, ethnic, race, gender and religious group to foster social order, whereas offensive language is often intended to directly insult a single individual [141]. Davidson et al. [25] collected a labelled dataset for identifying three categories of text classification: hate speech, offensive language and neither. The authors used, one-versus-rest multi-class classifier that transforms multi-labelled data into one binary classification task per class. The classifier was able to accurately detect both offensive language and neither at (91% and 95%) respectively. Yet, it did not perform well in hate speech (61%). Similarly, Warner and Hirschberg [138] presented an approach that relies on the tendency of using stereotypical words to capture hate speech. The stereotypical context may not necessary include abusive text, but refers to particular ethnic or minority group (e.g., Anti-African American text links to unemployment or childcare). The authors report that the classifier is able to detect hate speech on uni-gram textual features at 94% accuracy.

Sintsova and Pu [142] developed an approach to construct fine-grained emotion classifiers in replacement of data labelling utilising distant supervision. The method achieved high score of classifiers performance when testing hashtag-based text. Founta et al. [139] presented a deep learning interleaved approach to merge four features to discover the similarities and differences of behaviours among users. Text, content, user and network features, to detect sarcasm, cyberbullying, offensive language and hate speech, abusive text; reaching accuracy: 97%, 92%, 90%, 87%, 84% respectively. Similarly, Badjatiya et al. [143] attempted to investigate the methods of capturing hate speech using deep learning on various classifiers. The findings suggested that neural network frameworks are able to perform higher when incorporating gradient boosted decision trees. Gambäck and Sikdar [144] presented an algorithm for capturing hate speech utilising deep learning methods. In light of the fact that there are always publicly available, investigators will in general assess these frameworks using their own collected datasets via annotation tasks [145]. This makes direct correlations more channelling task and perhaps annotation becomes less reliable to use for further research.

Dinakar et al. [140] investigated the issue of detecting comments of online cyberbullying to compare the performance between multi-class and binary text classification

tasks. The comments cover sexuality, intelligence and race related topics. They conclude that JRip classifier is able to accurately detect sexuality (80.2%) followed by intelligence (70.4%) and race (68.3%). Zhao et al. [146] presented a machine learning approach for identifying the nearness of cyberbullying. Their methodology depends on word2vec and an extended list of predefined insulting terms on word embeddings. They set multiple weights on words to get cyberbullying features, which are then connected with enhanced Bag-of-Words (BOW) and latent semantic features to train SVM classifier. Nevertheless, the proposed enhanced BOW method overcame the traditional BOW model and showed improvement in detection performance reaching F1 score at 78%. Da Silva et al. [147] proposed an approach that investigates sentiment on tweets text that utilises a BOW model and feature hashing. Their analyses on tweet sentiments indicated that classifier ensembles can provide higher accuracy rate of classification performance.

Prior work continued to propose methods that encapsulate abusive content. For instance, email spam filtering has been an issue for multiple email platform providers and supervised or unsupervised learning methods have been recommended as a potential solution [148, 149, 150]. Recent work has looked at Natural Language Processing (NLP) approaches to overcome text classification tasks. For example, Schmidt and Wiegand [151] reviewed research on the automated capturing of hate speech utilising NLP approaches. The authors concluded that examining more linguistic features, e.g., politeness is a key factor for understanding the existence of hate speech. Also, heuristic data of users can help identify and understand the causes of abusive behaviour.

Danescu-Niculescu-Mizil et al. [152] introduced a computational framework through human annotation of data to detect textual features of politeness for examining social factors. The data was collected from Wikipedia and Stack Exchange platforms. The authors report that there is a negative correlation between politeness and social power. In particular, wikipedians are more likely to express politeness when they are elected, and become less polite after the election is over. Furthermore, users who are top-rated in the Stack Exchange platform are less likely to be polite, and those who are in lower-rate show more politeness. The framework later was expanded and publicly available [153] to all researchers to use. The three key categories of building textual-based detection of user behaviour are: polite, abusive and sentiment features as described in

Category	Feature Name	Description	Example
Polite (7)	Apologies	Remorseful affirmation	"I'm sorry for being so blun"
	Gratitude	Nature of being grateful.	"Thanks for your interest"
	Reasoning	Explicit reference to reasons	"I want to explain my offer price"
	Reassurance	Minimizing other's problems	"Don't worry, we're still on track"
	Hedges	Indicators of uncertainty	"I might take the deal"
	Negation	Contradiction words	"This cannot be your best offer"
	Questions	Question mark count	"Is this for real?"
Abusive (3)	Profanity	Vulgarity of all sorts	"The dang price is too high"
	Offensive	Direct insult	"U so retarded"
	Hate	Harmful intention to a group	"I'm going to blame the black man"
Emotional (3)	Positive	Positive emotion words	"that is a great deal"
	Negative	Negative emotion words	"that is a bad deal"
	Subjectivity	Personal opinion	"very great"

Table 2.3: Textual features of politeness [153], abusive [25] and sentiment [155].

Table 2.3. Locher [154] argued that people who are over-polite are more likely to be described as negative behaviour. Disagreement, on the other hand, can be a favoured reaction in the social interaction when an argument is raised and rivals are relied upon to their prospective peers. There are considerable cases that may affect online communities' expectations. In particular, social norms, conversational settings, discourse circumstances, users' age, status, or sexual orientation [154]. Nevertheless, polite and abusive disagreement has not been clearly investigated.

2.5.2 Activity-based detection

The presence of online behaviour is regularly based on activity which reflects the strength, persistence, correspondence and emotion of users perspectives. In particular, users who contribute with high reputation, correspondence and creativity, positive emotion and are concentrating around supporting and adding to the social network sites are portrayed as arbitrators [156].

Strijbos and De Laat [157] proposed a conceptual framework that aims to analyse interactions between peers. The framework suggests that role-taking tend to enhance the frequency of contribution goals in collaborative-learning environments, less engaged students were instead interested in individuals' goals since it requires less responsibilities. Golder and Donath [158] investigated multiple newsgroups on Usenet to un-

derstand the factors of social roles and interactions via an observational study. The findings suggested that users' behaviours affect the social roles. Specifically, newbies, tend to ask questions and help others most frequently. Slackers, tend to read messages, but take no action or answer a question. Chan et al. [159] attempted to cluster users roles on online forums by conducting an empirical analysis. There were nine features presented to characterise the roles of users' behaviours. The authors concluded that online communication can be characterised or grouped by behavioural roles.

On Twitter, Chatzakou et al. [160] explored three factors of identifying bullying and aggression behaviours online: user, network and textual factors. The authors reported that bullies tend to post hashtags and URLs most frequently than ordinary users, and have less connected friends or followers in the network. The main difference between aggression and bullying behaviour is that aggression is correlated to particular event or problem, whereas bullying behaviour is sequential and not accidental act with intention to cause harm and abuse the power against fellow peers [161]. Chatzakou et al. reported that users who commit aggression behaviour had their account suspended immediately. Bullies, on the other hand, did not receive any suspension, yet perform deletion of their accounts occasionally. Lozano et al. [162] investigated a Twitter group to capture racism against US Presidential Campaign in 2016. The authors reported that clustering users by *homophilous* behaviour can determine racist a negative tweets. This means that people are more likely to adjust their behaviour in accordance with peers who share similar interests.

Meire et al. [163] investigated the sentiment analysis of status posts on Facebook over time between Lagging and Leading variables, meaning before and after collecting posts. Then studied the correlation between sentiments and estimators variables. They utilised two classifiers: Random Forest and SVM evaluated on multiple two-fold cross-validation tests along with the Friedman test. The authors reported that estimators variables such as number of: likes, negative words and upper-case are significant variables for predicting sentiment text. Also, higher number of comments leads to absolute negative comments, and higher number of likes and upper-case most likely to post positive comments. The study findings suggested that considering Leading and Lagging data for the sentiment analysis can boost the performance of the classifier.

The emotion of the user contribution has additionally been utilised to identify the neg-

ative behaviour. Strijbos and De Laat [157], argued that trolls potential objective is to raise conversations on the subject of their enthusiasm for some close to home objectives. Cheng et al. [27] examined more than thirty-eight million comments of more than one million users from three news discussion communities to cluster users into two main types of groups based on banning activity (never/future). Composing in an unexpected way, future banned users vary from never banned users in their action. In particular, future-banned would post low quality content and receive aggressive reactions from the community members more than user who were never-banned. Another work likewise distinguished post recurrence as a sign of a low quality conversation [164].

Despite the fact that there is no generally concurred set of personal behaviour standards and marks, the social and specialised highlights considered by the above work while sorting behaviour do share a few motives including social interaction, information, entertainment and identification. Yet, users are most likely to adjust to a social network sites under scrutiny.

Cheng et al. [27] presented six categorises that can help capture anti-social behaviour. The categorises represents user, moderator and peer activities (e.g., posting, deletion, ratio of up-votes). Also, applied linguistics analysis such as readability and LIWC features. The classifier that used all behavioural sets achieved accuracy of 78% when determining whether a user will be banned or not. Another methodology [165] utilises Markov chain to figure the normal emanation probabilities of the n-grams in a user explicit discussion when the focused on abusive messages. The method looks at multiple contextual and behavioural aspects, e.g., (message or word length, number of bad words).

2.5.3 Assistive approaches

Chandrasekharan et al. [30] presented a moderation system that employs actions of moderators and reactions of users on ten online communities using cross-community learning known as Crossmod. The approach utilises a large corpus of pre-moderated comments with details about the actions of interventions. The data was collected from eleven active moderators on Reddit platform. Each action is associated with an agree-

ment score among moderators which allows the Crossmod system to make action decision based community norms. The system achieved an acceptable accuracy rate (86%) for capturing comments that are more likely to be removed by moderators. Furthermore, the moderators reported that the about 95% of detected comments were intended to be removed by moderators.

Chandrasekharan et al. [29] introduced the bag-of-community framework that computes post similarities of pre-moderated comments from Reddit and 4chan platforms to combat abusive behaviour of users. The approach reached high accuracy rate (91%) after examining more than 100K moderated comments. Cheng et al. [31] studied the impacts of both community reactions and users activity in four online communities. The reported analyses indicate that users who receive negative votes are more likely to evaluate their peers negatively, post most frequently and submit low quality of content.

2.5.4 Research gap

This section reviewed research in numerous interesting aspects and effectiveness of the interdisciplinary themes between social science and computing that can apply principles of computational methods such as NLP, machine learning, and empirical analysis of big data to understand the social complexity across online discussion communities and reduce causes of abusive behaviour. The approaches serve detecting variety of problems from email spam filtering to trolling and cyberbullying behaviours. Accordingly, regardless of whether a statement is considered as polite or abusive, or fitting or unseemly, relies to a great extent upon the norms of the similar context. Yet, we want to learn how/why text is different in online and in-person conversations, and if both settings can offer better textual features for detecting abusive content. Lastly, we want to present a disagreement detection model that can distinguish between polite and abusive disagreement.

2.6 Summary

This chapter reviewed the literature that assesses the primary characteristics and definitions of being abusive in online discussions. These definitions facilitate in under-

standing the characteristics of online abusive behaviours and motives to build a proper mechanism that combat these issues. In particular, this involves looking at nature of conversational settings, disagreement and perceptions of users and moderators. This thesis, in contrast, examines the correlations between behaviour on online and in-person, specifically, peer-group collaboration in affecting conversational rhythm as discussed in chapter 4, then show how disagreement can lead to abusive content in chapter 5. Chapter 3 addresses the fundamental expectations of content moderation. In addition, the dissertation expects to address many of the discussed above, however not all, of these difficulties. Specifically, conversations based on selected textual features for context analysis as shown in Table 2.3. These features are: polite, abusive and emotional content that can aid in finding issues with characteristics of online abuse in disagreement and disinhibition. Particularly, examining the differences between in-person and online conversations in 4 and investigating the spectrum of an argument from polite to abusive disagreement in 5. Disagreement is commonly described as a face-undermining act and negates an argument that may result in paying little mind to the degree of its deviant behaviour [166, 167]. In any case, contradiction is not generally uncountable in online discussion threads, which are a spot for individuals to openly communicate their fellow-peers, share thoughts, and express sentiments toward a specific issue.

Chapter 3

Moderators vs Contributors: The Case of Perspectives on Content Moderation

3.1 Introduction

As reported in the state-of-the-art work from literature in chapter 2, moderators' reactions create conflicts with contributors. In particular, the perceptions of moderation can often be linked to decision of intervention. Nevertheless, moderators' reactions are not always the same, i.e., some moderators may become more tolerant about particular post/comment, yet others may not take similar action. This chapter seeks to investigate the perceptions of moderation activities including strategies and reasons for handling interventions. The goal of this chapter is to identify the perceptions of moderators and gaps between moderators and contributors on content moderation.

Content moderation has become a de facto necessity on platforms that enable user contributed content, such as Reddit, Facebook, Twitter. Such sites enable users to contribute their own content as desired, including text, images and video, enabling global platforms for civic discourse [79]. However, the growing popularity of such sites inevitably draws the attention of regulators and requires platform providers to develop policies describing acceptable standards for contributed content, reflecting legal con-

straints and societal norms. Policies may address acceptable forms of conduct in discussion (avoidance of insulting or ad hominem messages directed at other others) or limit what topics may be acceptability portrayed or discussed.

On open platforms, such policies require enforcement, as users may not always be willing to comply with, or be aware of the necessary standards. Many platforms have adopted the practice of employing *moderators*, privileged users with additional capabilities to restrict or ban content contributed by others for this purpose. However, the scale of the content generated on social media platforms has grown dramatically in recent years. According to the Internet Live Stats service¹, Twitter users generate 200 billion tweets per year and according to Reddit's review of 2018, users made 1.2 billion comments on 153 million posts². The sheer size of content has led platforms to seek opportunities to scale the moderation process itself. Prior work on content moderation has focused research on filtering or pre-labelled text by leveraging crowd-sourcing and utilising natural language processing approaches [168, 85]. As a consequence of these developments, moderation activities may be fully automated, human with automated assistance or fully managed by the human moderator. This infrastructure creates a complex eco-system of influences on moderation decisions. Policies may be fully embedded in the design and implementation of filtering algorithms. More commonly, filtering mechanisms may be used to guide moderators and prioritise their decision making. This interplay between social and technical components of the moderation infrastructure could have a more subtle influence on the behaviour and practice of moderators and their consequent decisions.

Consequently, human decision making remains a key component of moderation activity. However, relatively little research has been conducted to understand the perspectives, attitudes or decision making processes of moderators on social media platforms. Previous work has investigated the impact of moderation on contributors, for example [70, 72]. However, there is limited research that seeks to understand the moderation process from the perspective of the instigators, i.e. the moderators themselves. This is important because the design of both facilities for moderation and assistive technologies for filtering content could be significantly impacted by the motivations,

¹<https://www.internetlivestats.com/twitter-statistics/>

²<https://redditblog.com/2018/12/04/reddit-year-in-review-2018/>

expectations and practices of their users, the moderators.

Therefore, this chapter seeks to investigate these issues through a survey of moderators on social media. To investigate this phenomenon, the following exploratory research questions were developed to guide this thesis:

RQ1. What role do moderators perceive for themselves on Reddit discussions?

RQ2. When, how and why do moderators intervene on discussions on Reddit?

The above questions highlight three main themes of content moderation: intervention activity, motivation and role expectation. These themes emerged within the literature. To investigate these questions, we used a mixed approach of qualitative and quantitative analysis. To begin, the study surveyed (N = 218) moderators from Reddit platform to investigate the related issues on content moderation. The survey questions were designed to understand the demographics of moderators on Reddit, their views on their role as moderators within a wider community of users, the effort contributed to undertaking moderation activities and the actions they take when moderating discussions, including removing content and banning user accounts. The survey was complemented by an analysis of actual user and moderator activity on Reddit, using a recent sample of data on comments. This second source of evidence provided an understanding of actual work patterns of both moderators and the wider Reddit community of users.

The key findings from the study are that:

- Moderators on Reddit are largely *community* motivated and have a strong sense of the importance of fostering a sense of community amongst their contributors and users. Despite this, a significant minority of moderators on Reddit undertook moderation activities on behalf of an employee.
- Although moderators can and do take punitive action with significant autonomy, they also feel the need to engage with contributors and help them to improve the quality of their contributions where possible.
- Frequency of reviews amongst moderators per week varies considerably, but the intervention rate declines as the total number of reviews performed increases.

- Moderator and contributor peak activity times are divergent. Contributors are most active during working hours. Moderators are most active at weekends.

This chapter is structured as follows. The first section reviews forms of moderation described in the literature. Section 3.4 describes the design of the survey and complementary data gathering of contributor and moderator activity. Section 3.5 summarises the findings from the exploratory survey and 3.6 discusses the limitations, then 3.7 concludes by addressing the key summaries of the work.

3.2 Forms of moderation

Several factors account for the nature of moderation in different settings.

Demographics. There has been research conducted on age and region (origin county) on user's perspectives of content moderation [79]. Hurt et al. [169] suggests that younger generation at collage-level are more likely to be active and express emotions on an academic discussion on social platform like Facebook. Alkharashi et al. [170] shows preliminary analysis of how gender and background play significant role in shaping opinion about what is (Un)acceptable to discuss on group discussions in online compared with in-person settings. How these influences on viewers and contributors to content influence the practice of moderation is less clear.

Sources of policies. Caplan[171] identified three approaches to the development of policies of moderation on different platforms. Artisanal approaches in which the platform's in-house employees (largely manually) review content on a case by case basis. Community driven approaches, in contrast, rely on the development and enforcement of standards within the contributor community itself, including the collective appointment of enforcing moderators. Finally, industrial moderation refers to the development of formal rules that can be enforced by a large workforce of paid employees with little recourse to flexibility or discretion.

Pre vs. post moderation. Moderation may also vary depending on where is it applied. Pre-moderation requires the platform owner to assign moderators that can review and approve the content before it is published for consumption by others. This is mostly

used in product reviews or multimedia post including images, videos or audio files. While this approach preserves the quality, it leads contributors to be less proactive [172]. Post moderation is a synchronous intervention that can be reviewed and approved after submission. This approach seems to be popular in discussion communities to keep the velocity of interaction among users, yet experiences a higher risk of trolling and abusive activities [27, 29].

Community Feedback. This kind of moderation works on the policy that promotes methods to flag or removes unwelcome content and reported by the users. That is why this is not proper for highly populated and conscious platforms where online users are not as scrupulous and productive. Another way is to build a rating system (e.g. votes or likes) to allow users to provide their feedback on posted content. Although this is a desirable and influential method of productivity, it still suffers much in the quality of content due to complicated social feedback effects [72]. Additionally, this relies on a passive moderation process that consumes a remarkable amount of time to detect abusive content.

Technology assistance. Extensive research on automated moderation approaches to supervising online that detect abusive content or political misinformation within textual features or diffusion network [135, 173] and unusual or unhealthy behavioural patterns [174, 175] can be identified based on topic modelling and human annotation showing the appearance of offences by user behaviour. Yet, the literature in online moderation requires empirical studies to understand content moderation strategies, motivates, and roles.

3.3 The Reddit platform

The reason for selecting the Reddit platform as a focus of study is due to its popularity for hosting discussions across a wide variety of topics. Reddit is organised into *subreddits*, each covering a particular topic, for example, r/politics, r/ukpolitics and r/cinema. As of 2018, Reddit has more than 1.2 million subreddits, of which 140,000 are considered active. Users of the platform may be passive readers, active contributors of content, or moderators with responsibility for removing content or imposing other sanctions on users.

Contributions to a subreddit can be categorised as either posts or comments. Posts are the original source of a topic for a conversation, including, for example, links to news stories on other platforms. Comments can be either added to these posts or to earlier comments, creating threads of conversation.

Contributors may also vote up (positively) or down (negatively) on either a post or a comment. The Reddit platform automatically calculates two metrics for each voted contribution: score and controversiality. The score of a post is its net positive rating (up - down votes). Controversiality provides a reflection of how contentious a contribution is with other users on the platform. A magnitude is calculated as the total number of up and down votes. The ratio of up votes to down votes, or down votes to up votes, depending on which is the smaller is then applied as an exponent, i.e: $\text{controversial} = \text{magnitude}^{\text{balance}}$ where $\text{magnitude} = \text{up} + \text{down}$ and $\text{balance} = \text{up}/\text{down}$ if $\text{up} < \text{down}$ else up/down . Therefore, contributions with high total votes (both up and down) and with similar up and down votes are defined to be controversial.

Caplan [171] defines Reddit as operating a *community reliant* content moderation platform, reliant on volunteer moderators to review and if necessary intervene on contributions. Moderators may themselves be contributors and may both contribute to and moderate more than a single subreddit. Depending on their privileges, moderators may be able to edit or remove contributions, change the access rights of other users or flag content with warnings.

3.4 Method

To investigate the research questions, a survey of moderators on Reddit was undertaken, seeking to understand their perspectives on moderator practice. This survey was complemented with a large scale analysis of moderator actions to remove content from discussion threads using publicly available data. This Section describes the two data gathering methods separately.

3.4.1 Survey of Moderators

The first step was to design a survey of content moderators on Reddit, in order to understand the activity from their perspective.

Survey Design: Besides questions concerning consent, follow up and demographics, the survey consisted of 7 questions addressing moderation activity (questions 5-8), characterisation of interventions (questions 9 and 10) and the role of the moderator (question 11). The questions asked are summarised in Table 3.1. The questions were developed within the interdisciplinary research team iteratively between social science and computing science scholars based on discussions and pilot experiments. Also, the questions targeted three main issues that occurred in literature as research gaps. These questions provide a rigorous learning about the behaviour of moderators and contributors in the online communities. The first theme has Q5-Q8 which aims to investigate the activity of moderators when reviewing and removing a comment on Reddit. This includes the duration time per week. Also, includes the type of action to remove a comment. The following theme (Q9-Q10) aims to understand the rationale behind removing a comment and what type of community that requires the most of intervention. The final theme which is Q11 aims to investigate the perception of moderator role. In particular, see how moderators describe/view their duties of each community. The survey design and plan received ethical approval (Application No. 300180287).

Recruitment of participants: The survey was open from July 2019 to October 2019. Participants were recruited opportunistically, using three different methods. It began by using the subreddit List service³ to identify 500 of the most popular subreddits and then recorded the unique identifier for every moderator listed. Each subreddit contains at least one or two moderators which may moderate more than one subreddit. There are two approaches for contacting moderators, either by sending message to the entire group of mods listed in the subreddit homepage or sending direct message to each individual mod in each subreddit. Firstly, a message was sent to the moderator group for the subreddit, explaining the reason for contacting them and inviting them to participate in the survey. The total number of moderators is 5094.

³<http://redditlist.com>

(A) Intervention activity

Q5. In an average seven-day week, approximately how much time do you spend on reviewing content for moderation? (One hour/ Half a day/One day/ Two to three full days/ Five or more full days)

Q6. In an average seven day week, how many posts would you review for potential moderation?

Q7. In an average seven-day week, how many posts do you intervene on?

Q8. What interventions might you perform? (Deleting a post/ Suspending user account/ Permanently banning user account/ Escalation to higher authority for for review or second opinion/ Other – please specify)

(B) Intervention motivation

Q9. What is your most common reason for intervening on a post?

Q10. Within the forums you moderate, what topics require the most frequent intervention?

(C) Moderator role

Q11. Which of the following describes your role as a moderator (select all that apply)? (A member of a community, responsible for preserving community values/An independent moderator, guided by personal values/Other – please specify)

Table 3.1: The survey design highlighting three key moderation themes to study the moderator’s behaviour and perception. The questionnaire design leverages a mixed of short and open-ended questions in some cases.

Later, to strengthen the sample size, a direct message via Reddit was sent to each individual moderator, again inviting them to participate in the survey. Following this strategy, a direct contact of 250 moderators was taken from the 200 subreddits. Finally, several participants responded suggesting to post the survey link to some specific subreddits `r/sample size` and `r/needamod`, which was followed based on the recommendations.

3.4.2 Collection of Moderator Action Data

To complement the results of the survey, data was collected on the actions taken by moderators to remove content from subreddits. To do this, the Google BigQuery service was used to retrieve a sample of comments, including removed comments, posted during March 2019. In total, this data comprises 168 million comments, of which 5% were removed by a moderator and 3% were removed by the user. To make use of this data, several queries were performed on the BigQuery service to provide us with a

summary of behaviour over the four week period considering the following metrics:

- *User activity.* The empirical analysis of moderator behaviour began by considering the activity of their subjects: the users of the platform. To understand when Reddit users are most active, the total number of contributions made in each one hour time slot during the sample period were calculated.
- *Controversiality of comments.* The Reddit platform automatically calculates a contribution score for each contribution, based on the relationship between the total number of up and down votes. Average controversiality was calculated for each one hour time period during the sample period.
- *Moderator activity.* There are multiple actions that a moderator can do based on their account privileges. However, these are not accessible to the public and therefore it is not possible to get all actions listed on the moderation log unless a permission is granted to moderate a subreddit. Therefore, to approximate when moderators are most active on Reddit, the total number of comments that were removed by moderators for each hour over the four week period were calculated. In some cases, the removal is automated by the NLP training decisions, where the intervention is immediate and will not be archived.

These three metrics were used to understand when user activity occurs on Reddit, when this activity is most controversial and when moderators are most likely to be active in removing content.

3.5 Findings

A total number of 220 responses were received. Two responses were excluded due to offensive or irrelevant responses to questions, leaving a total sample size of N=218. In some cases, users did not provide complete answers to all questions. Where this occurs the reduced sample size for the purposes of the analysis is stated. Where possible, the responses also illustrate some of the analysis with free text answers provided by the respondents.

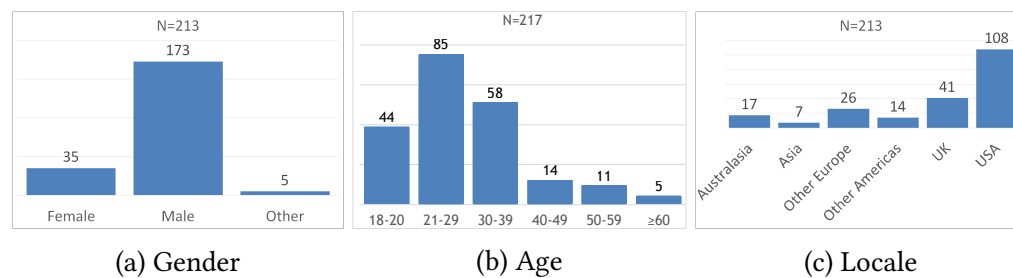


Figure 3.1: Summary of demographics ($N = 218$) recruited from reddit in accordance with responses collected from the survey. The different sample number in each figure reflects the total number of positive responses, i.e., some respondents also chose to not reveal their gender, age and nationality.

3.5.1 Demographics

The summary of demographics is shown in Figure 3.1 and in Table 3.2.

Gender: The gender breakdown of the respondents is shown in Figure 3.1a, divided between Female (16.1%), Male (79.9%) and Other (1.4%). A small number of participants (2.8%) preferred not to state their gender.

Age: 39.2% of the respondents are aged between 21-29, followed by 30-39 with 26.7% and 20.3% for 18-20. The remaining three age groups are 6.5%, 5.1% and 2.3% respectively.

Locale: Most of the respondents are from North America (56%) and Europe (32%). A small number were from Australasia (8%) and Asia (3%). Three respondents did not give their country location, with one participant noting that this omission was owing to having received death threats.

These figures are broadly comparable with those for Reddit Users, according to a survey by Pew Research [176], which found that redditors were predominantly young, male and based in North America. This suggests that the profile of moderators on Reddit is similar to the profile of contributors (Redditors).

3.5.2 Moderator roles

Before considering moderator behaviour, the study contextualises the analysis by considering Q11, how moderators view their roles. Figure 3.2 illustrates the different re-

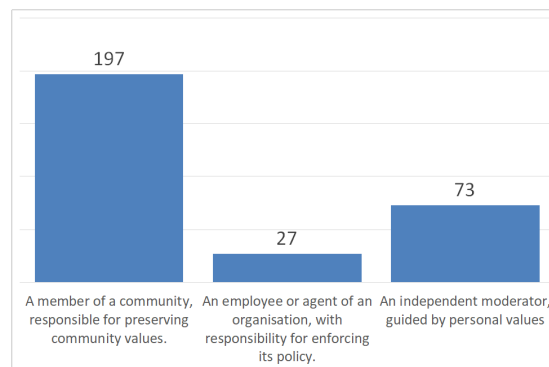


Figure 3.2: Moderator views on their roles.

sponses selected by moderators for this question. Note that multiple answers to this question were permitted. Almost all respondents (197, 90%) reported that they considered themselves to be a member of the community that they moderate, and were responsible for upholding the community’s values. Conversely, just 27 (12.3%) reported that they were an employee of an organisation responsible for enforcing that organisation’s policies. This suggests that the respondents were overwhelmingly volunteers and would therefore undertake the moderation activity outside of their professional roles.

Several of the respondents illustrated this choice with additional commentary. For example, M168 stated that:

“Like a democratic republic our forum has been established with certain guiding principles. Like fundamental human rights they must be held as inviolate as possible. The community standards or preferences might sometimes conflict with those principles. In that case I act to preserve them unless they are specifically called out for modification. That can sometimes put me in conflict with the community.” (Male, 50-59, USA).

The observation about the negotiability of the rules within the community and the potential for conflict highlights the role of the moderator as a “good citizen” member of the wider community. Another perspective on the reason for selecting more than one role is described by M31:

“Subs have historical social norms. Rules are often codified norms. At other times those norms are not codified, but carry on as norms. Each rule was made to make 'advice giving and asking' easier. Or they were introduced because of an event(s) which were seen as detrimental by the, then, mod team. Everyone, as individuals, have personal values. We're human. Not robots. Internal debate is encouraged and we actively try to make our community a better place. We've active. Not passive". (Male, 21-29, UK).

Another view by M155 highlighted the importance of the shared values of moderators.

I'm independent in that I have complete control and make my own judgments, but I'm also a member of a team of people who are all independent in the same way, and who have all been chosen because we share common personal values for the community. (Female, 21-29, Australia)

These findings suggest the potential for platforms to better leverage the sense of belonging and collective autonomy amongst their moderators and the role they play within the wider community.

3.5.3 Working patterns

Figure 3.3 provides two perspectives on the amount of effort moderators contribute to reviewing contributions on Reddit. Figure 3.3a illustrates the amount of time a moderator contributes to reviewing contributions over a working week. The figure illustrates that very few moderators spend more than a full day reviewing contributions each week (12.4%), whilst almost half of respondents (46%) spend half a day and almost a quarter (just 23.5%) an hour each week. This highlights the fact that moderators are volunteers, working in their spare time to enhance the quality of discussion within their community of users.

Separately, Figure 3.3b presents a scatter plot of reviews against actual interventions (decisions to take action on content) by moderators for up to 1500 reviews. The chart illustrates several trends. First, the chart shows that the majority of moderators review

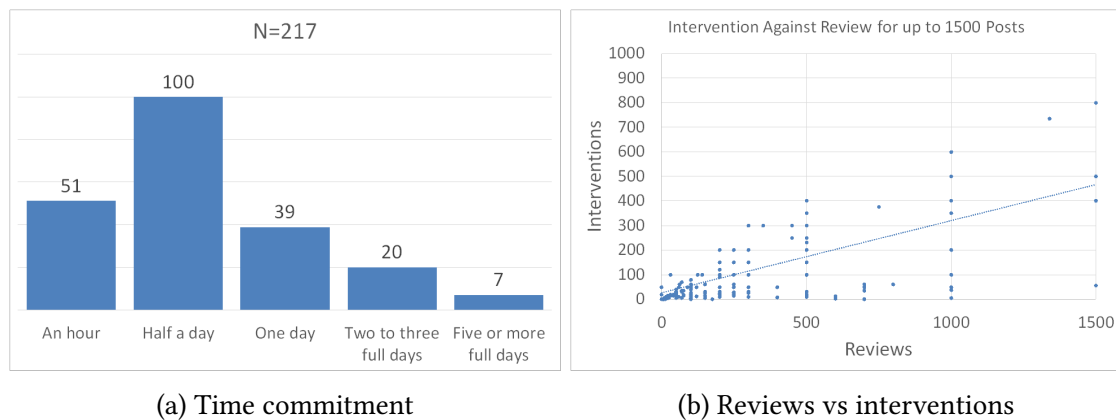
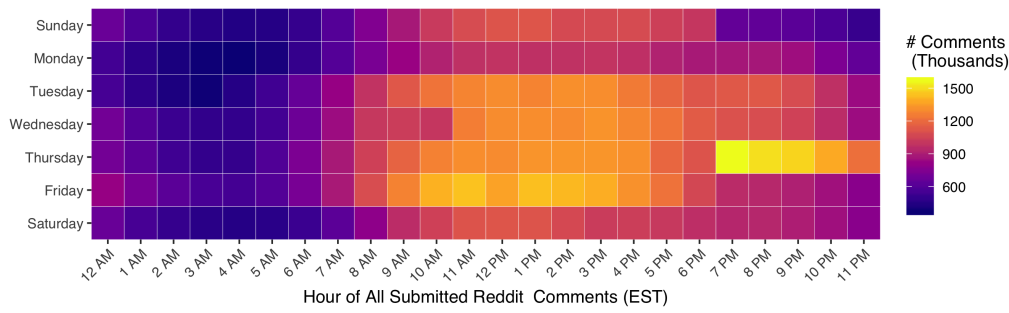


Figure 3.3: Time commitment and reviews undertaken by moderators during a seven day week

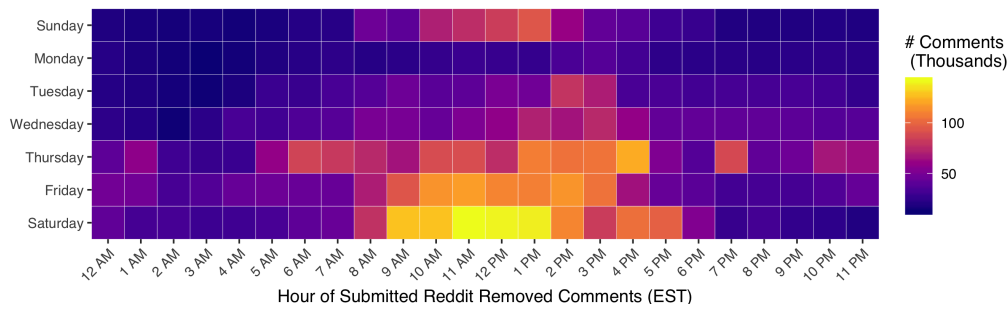
less than 250 contributions per week, with 83.0% reviewing less than 1000 contributions per week. Despite this there are a number of outliers with a small number of moderators claiming to review many thousands of contributions per week. Caution must be adopted here in relation to these responses, since it is possible that the question may have been misunderstood. Respondents may have considered the question to concern total reviews over all time. Alternatively, the respondents may be counting all aggregating all contributions as having been reviewed when a single post is moderated along with all sub-comments.

A second trend found is that the maximum intervention rate declines as review rate increases. This suggests that reviewers of larger numbers of contributions are more willing to give contributors the benefit of the doubt, rather than take a precautionary approach. Also, investigated whether gender had any impact upon review rates. However, there was no obvious effect of gender. Once the outliers described above were excluded female moderators intervened on approximately 40% of contributions and male moderators intervened on 31.1%. However, this was not a statistically significant difference (p -value: 0.5695) using regression analysis for gender against the two independent variables interventions and reviews.

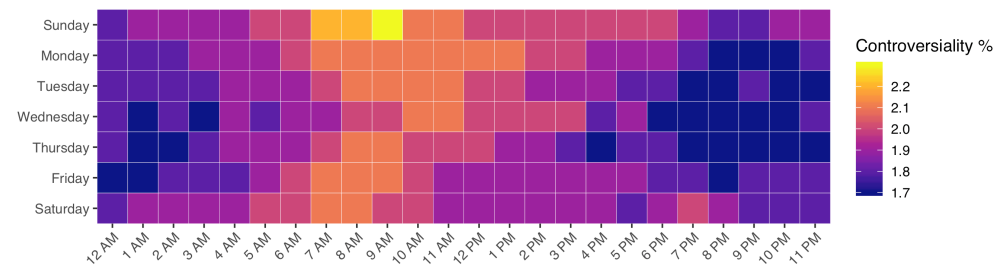
As described above, the survey questions of moderator activity were complemented with empirical analysis of contributions and moderations of Reddit during the month of March 2019, using the Google BigQuery service and Reddit contribution dataset.



(a) Total contributions submitted per hour.



(b) Removed comments per hour



(c) Controversiality of comment

Figure 3.4: Heat maps of activity on Reddit per hour over four working weeks in March 2019. Figure (a) shows that contributors tend to be more active during evening time and weekends. Figure (b) shows that moderators are more active during working hours between 8 am to 5 pm. Moderators tend to be more active at weekends, whilst contributors are more active during the week. Figure (c) shows that comments tend to receive controversial votes from users mostly during morning time.

Figure 3.4 illustrates the results of this analysis. The figure plots each metric in week-hourly time slots over the the whole of March 2019 (scores for Friday, Saturday and Sunday are weighted). Since of the number of days on March in 2019 is 31, there were five Fridays, Saturdays and Sundays, rather than four. So, we combined the additional week to the rest of the month and divided by four weeks.

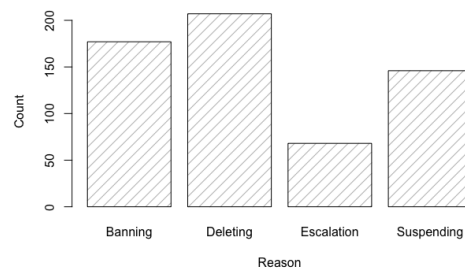


Figure 3.5: Actions taken to intervene on posts.

Several trends are apparent from the heat maps. First, contribution activity overall peak during working hours. The sudden increase in contributions in the morning of each day and more gradual decline in evenings may reflect the distribution of working hours across Eastern, Central and Pacific time zones in the USA. The peak in activity for Thursday may be due to the fatal attack on a mosque in New Zealand on the 15th March 2019, and the consequent reaction on various social media platforms.

Second, by contrast, controversial contributions tend to be spread throughout the whole week, with greater concentration of controversial contributions in the morning. Note that the graph shows the percentage of controversial material by hour, so this is relative to total contributions. This suggests that more antagonistic discussion occurs throughout the week and in particular in the early morning.

Finally, the chart concerning removed content suggests that moderators are most active at the weekends in contrast to the time when contributors are most active, or when more controversial contributions are made. This striking difference between working patterns of contributors and moderators may have implications for the dissemination of inappropriate material, since moderation may be less stringent during times when controversial contributors are most active.

3.5.4 Intervention actions

Next, the actions that respondents took in order to intervene on a contribution and the reasons for doing so were considered. Figure 3.5 shows the actions taken by the respondents when intervening on a post, and Table 3.2 shows the actions by background factor. By far the most popular response by almost all respondents was simply

Background		Prop.	Remove	Suspend	Ban	Escalate
Gender	Female	(16%)	0.94	0.77	0.77	0.20
	Male	(79%)	0.95	0.66	0.84	0.32
Age	18-20	(20%)	0.95	0.66	0.82	0.36
	21-29	(39%)	0.94	0.62	0.88	0.33
	30-39	(27%)	0.95	0.69	0.74	0.26
	40-49	(6%)	0.93	0.86	0.86	0.21
	50-59	(5%)	1.00	0.64	0.64	0.55
	≥ 60	(2%)	1.00	0.80	0.80	N/A
Locale	Asia	(3%)	1.00	0.57	1.00	N/A
	N. America	(57%)	0.94	0.74	0.80	0.26
	Europe	(31%)	0.94	0.50	0.79	0.38
	Oceania	(8%)	1.00	0.82	0.94	0.53

Table 3.2: Summary of demographics ($N = 218$) recruited from reddit in accordance with responses collected from Q8 by three-dimensional background factors. The number besides bar chart is the percentage of total number of responses received from the respondents per group sample.

deleting a contribution. Less common responses included suspending and/or banning a user account, although it can be noted that more than two thirds of respondents reported one of these two actions. Conversely, less than a quarter (68) of respondents reported the escalation of cases to a higher authority for review.

This complements the finding above regarding the roles that moderators perceive for themselves within their community. Not only are moderators volunteers, but they consider themselves to have considerable autonomy within their discussion group to make decisions about the behaviours of others and impose sanctions, without recourse to higher authorities. In addition, the responses suggest that the respondents are comfortable in applying the most severe sanction (banning a user account) if necessary. However, further research is required to understand the relative frequencies of these actions and how this behaviour might vary between moderators.

About 26.3% have reported other kind of intervention actions including locking a thread or post, adjusting spam or moderator filters or warning a user via personal message (PM) that their contribution or conduct is inappropriate. Moderators also reported asking contributors to adjust their contributions before approving them. One moderator

commented on this interaction with contributors:

“A deletion is traditionally followed by a warning of some kind, which may or may not lead to discussions on the rule sets with individual users”
(Female, 18-20, USA)

This indicates again, that moderators see themselves as “good citizens” within their communities, striving to improve the quality of debate through negotiation with users, as much as guardians of the content.

3.5.5 Reasons for intervention

Answers to Question 10 were free text. The word cloud in Figure 3.6a illustrates the common response terms provided by respondents to this question. Responses were reviewed following the close of the survey in order to identify recurring themes. These were then coded as common terms that were then grouped. For example, rule breaking as responses that included the terms ‘rule’, ‘breaking’ or ‘violation’ were categorised. Figure 3.6b summarises the reasons reported by the respondents for intervening on a post. Note that a response might well fit into several categories, so these response categories are none exclusive.

The figure shows that the most common reason reported for intervention (93 respondents, 42.7%) was rule breaking. This aligns with the perceptions of moderators as guardians of community values, ensuring a higher standard of debate. Other responses provide more specific reasons for potential rule breaking. For example, insulting posts (40, 18.3%) and attempts to post spam or advertising rather than genuine content (37, 17.0%). One moderator described the sophisticated spamming strategies that they needed to work to eliminate:

“Probably spam, this is something that happens on all my subreddits. There are two major types of spammers that I come across. Ones where they are still building up karma to spam later or otherwise astroturf, and ones that have already been activated and spam sketchy t-shirt sites that steal your credit card info or whatever else.” (Male, 18-20, Estonia).

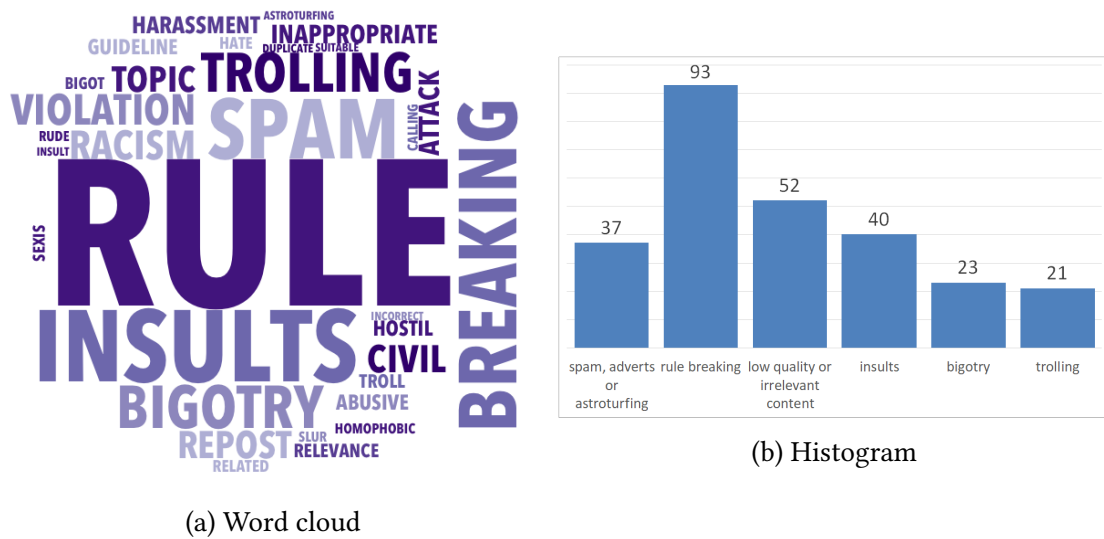


Figure 3.6: Categories of reasons for interventions and associated word cloud

This illustrates how spammers will attempt to exploit the *Karma* system of rewards on Reddit in order to gain a more trusted status that can then be later exploited. The comment also suggests that moderators are adopting strategies to anticipate and detect this behaviour.

More subtle aspects of quality control are also included in responses. Some moderators (52, 23.9%) report removing content that is genuine, but low quality, or irrelevant to the discussion. A small number of moderators also report removal of content due to *trolling*, [27], the use of social media to deliberately provoke or bait other contributors. For example, M158 reported that:

“If we have to intervene with back and forth commentary, the most common topic is flags/posters. Somebody will post a picture of their home gym (the broad topic of our forum), and somebody else comment on a flag or poster in the background. E.g., a flag that is intended to support law enforcement, or a U.S. flag that is hung incorrectly, or a modified version of the U.S. flag that is intended to represent a particular movement or protest will often spawn hateful comments” .(Male, 21-29, USA).

Another reason for intervening and removing content reported was the risk of privacy

breaches. The following case scenario is shared by M108:

“Identities should always be kept private, but we do often see people who don’t understand this. I usually remove the content and ban the original poster (OP) but occasionally I’ll encounter posts where users are encouraging the OP to divulge private info. In those cases, I remove any pertaining content and ban all users involved. There have been times where I report a case to Reddit’s administrators [...]” (Male, 21-29, USA).

This suggests again that moderators see themselves as good citizen protectors of the contributors to their community, as much as guardians of the content.

3.5.6 Topics for intervention

Figure 3.7 summarises the topics reported by respondents on which they most commonly have to intervene (Question 11). Again, answers to this question were free-text, so responses were categorised in a similar way to Question 10. A variety of topics were reported. Figure 3.7b shows that the most common topic to intervene on concerned politics (51, 23.4%). As one respondent commented:

“In general, when Donald Trump is an issue for debate, arguments get heated and insults become more common. Another big topic are ‘foreigners’, especially topics about Muslim faith. It doesn’t matter if it is about them directly, or politicians who are considered to be supportive of them.” (Male, 21-29, Germany).

Reviewing the free text answers, many of the responses were hard to categorise into a topic, as many respondents re-stated the reason for the intervention action, rather than the underlying topic that caused the intervention. As a consequence, it was unclear to identify significant themes in these responses. On the one hand, this may be because the moderators were surveyed across a wide variety of platforms. Another possibility is that the topic of discussion may be less important in stimulating moderator activity. Rather, the moderator intervenes to enhance or penalise contributor behaviour, independently of the focus of discussion. Further research is required to assess the importance of topic in stimulating moderation.

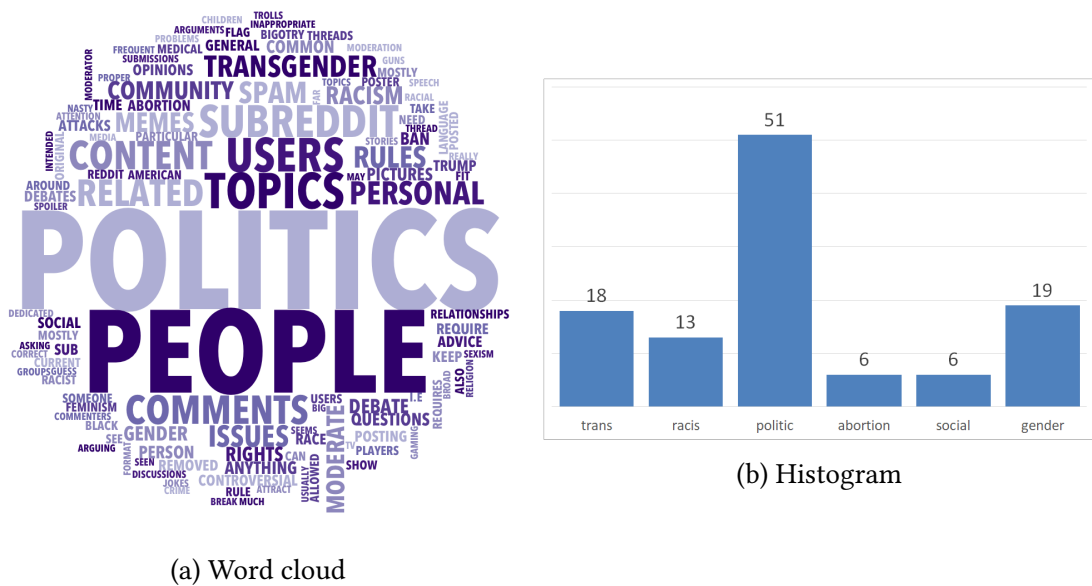


Figure 3.7: Categories of topics for interventions and associated word cloud

3.6 Discussion

The survey questions highlighted three significant themes relate to the perception of content intervention: actions, motivation and role of moderator. Given the results from the analysis, we may argue that social systems still suffer from optimizing the automated content moderation process due to the lack of understanding the structure of intervention perspectives against community rules or norms.

3.6.1 Implications for the development of moderation system

A key finding from the study is that moderators perceive themselves as custodians of their community and, as well as undertaking punitive action, work hard to engage in dialogue with their contributors to assist them in conforming with a subreddit’s policies. This finding has implications for the development of facilities for moderation, suggesting that moderators might benefit from more sophisticated copy-editing facilities, such as the ability to propose corrections to content on a post as part of a peer-review workflow.

Moreover, the activity times was discovered of contributors and moderators that are divergent. Contributions to Reddit are normally highest during working hours between Monday and Friday. Conversely, moderators are most active during weekends. This observation may have implications for the design of moderation systems as well. For example, contributions that violate policies may persist for longer on Reddit if they are published early in the working week. This may suggest that review filters may need to be adaptive to be more sensitive at these times, or prioritise cases for review based on the time they were posted. Also, moderators reported that topics that relate to political views, religion and gender are at higher risks of encountering abusive content. Mostly moderators intervene when contributors are breaking the community rules. Nevertheless, social system designers may consider developing divergent strategies for moderation requirements depending on the topic. For example, assign an assessment task to contributors to evaluate removed comment by a moderator from most intervened topics and reward contributors after competing the task successfully. This approach can enhance positive engagement among contributors and moderators to give wide perspective of reasons of intervention.

Informally, it was astonishing participation reported by the largely positive response from moderators to the invitation to participate in the survey and pleased with the high response rate. In some cases, respondents made helpful suggestions as to how the survey could be disseminated and were supportive of the research. In addition, moderators indicated that the survey was the first time they had been contacted by researchers to understand their perspective, strengthening the impression that this is and an under-investigated area of work on social media.

3.6.2 Limitations

The work presented in this chapter is a first exploratory study of the perspectives of moderators, and as such has some necessary limitations. First, the study was restricted to a single social media platform, embodying the *community driven* approach to policy development and enforcement [171]. This inevitably narrowed the perceptions of roles reported by moderators who responded to the survey, compared to those found on other platforms, such as corporate news sites or social networking sites such as

Facebook. It is unclear what impact this might have on the self-reported actions or autonomy of decision making. Nevertheless, few of moderators have reported that they described their role as an acting on behalf of an employee.

Second, this study adopted opportunistic convenience sampling, contacting moderators of popular subreddits and inviting them to participate. As a consequence, it is difficult to determine the extent to which the demographics of the moderator sample is reflective of the demographics of the population of moderators on Reddit. However, it was noted that the demographic is broadly comparable to that of users of Reddit as a whole.

Thirdly, the study deliberately chose to limit the scope of the survey questions to maximise the number of responses received. This necessarily limits the extent to which it can report on more nuanced aspects of moderator perspectives and behaviour. For example, it was not possible to determine the relative frequencies of different interventions (removal of a post versus banning an account for example). Nevertheless, the use of a small number of free text answers did provide valuable insights into the moderator's perspective on their role. For example, many of the moderators reported on the extent to which they used dialogue with (often new) users to enhance the quality of contributions.

3.7 Summary

This chapter has presented a first survey of the perspectives and activities of moderators on a social discussion platform, Reddit from 218 moderators that were invited from most active 500 subreddits. Three main themes were investigated including intervention activity, intervention motivation and role of moderators. The analysis suggests several trends and gaps between moderators and contributors that can impact the intervention activity and motivation.

Based on the findings that were reported in this chapter, there are several cases of abusive behaviour that stipulate further investigation. Thus, in the following chapter, a longitudinal study is conducted to uncover the differences between in-person and online peer-group conversations.

RQ1. What role do moderators perceive for themselves on Reddit discussions?

The answer to RQ1 is that the majority of moderators reported their role as member of community who is responsible for preserving online community values.

RQ2. When, how and why do moderators intervene on discussions on Reddit?

The answer to RQ2 is that intervention activities decreases when the number of reviewed posts is higher during working days and hours (9 am-5 pm ; Mon - Fri) mostly intervene in political discussions to remove comment or post mainly due to breaking the community rules.

Chapter 4

Online vs In-person Conversation: The Case of a Peer-Group Project in a Learning Environment

4.1 Introduction

The previous Chapter 3 as identified in the survey of moderators, has focused on the perceptions of moderators in Reddit and analysing contributors' activities, i.e., posting vs removal of comments. There are several forms of online abuse that were reported by moderators such as hate speech, offensive language, death threats or any action that violates community guidelines. It is well recognised in Chapter 2 and Chapter 3 that abusive behaviour is prevalent on online platforms and can have extremely serious consequences. For example, there was a mass shooting on two Islamic prayer centres, where a middle-aged person decided to broadcast this attack on Facebook and link it to a white supremacy group. At least ten minutes before the attack, the attacker posted a comment on /pol/ board at 8chan targeting Alt-right group to invite to his Facebook page and published a radical and hateful document [177].

However, little is understood as to the causes of this abusive behaviour in online discussions. Some guidelines on online conduct recommend that online behaviour should be the same as the UK home office states:

Your behaviour online and your behaviour in-person should be the same. Your online behaviour should reflect your in-person behaviour — you shouldn't behave differently simply because you're online [178, p. 6].

In contrast in this chapter, a direct comparison is made between contributions to equivalent discussion topics in online and in-person settings. One particular and significant problem is learning about user's behaviour inside and outside a community to be able to characterise abuse differences in how users pose an argument in a discussion in an online setting compared with an in-person setting. Most of the time, online abusive incidents can lead to in-person violence or aggression attack [179].

Existing research has been conducted in both settings, however, relatively little is known about the difference between online and in-person conversations that can cause abusive behaviour. Prior work has attempted to empirically verify causal connections to show the difference between online and in-person groups in a synchronous communication that is related to e-learning [180], peer support for children [181], social support [182], interviews [183], online gaming [184], and political engagement [185]. Extensive researches reported qualitative studies that focus on analysing communication consequences of online and in-person behaviour discussion in civic engagement [186], frequently by studying the behaviour of a significant number of users. In particular, areas such as criminology [187] where a study of low income people was conducted to understand the affect between online and in-person communications about crime conversations. A more comprehensive perception of abusive behaviour demands a quantitative, qualitative, longitudinal study by measuring the differences between conversations. This can guide new techniques for classifying unwanted comments and lessening vandalism behaviour, which can eventually produce stronger online communities. The motivated research statement can be expressed as *online conversations can disinhibit communication on collaborative environment and may lead to abusive behaviour.*

Several platforms use a variety of human and auto-moderation approaches that are intended to identify abusive behaviour on online discussions. These approaches including human moderation, community reaction with votes, and flagging or reporting

a suspicious post or comment. Nevertheless, it is unclear how and why both communication settings are the same or different.

In this chapter, the research is seeking to uncover the presence of stimulator of abuse and absence of guards against abuse in online conversations, compared to in-person data by examining several types of data (e.g., audio, survey, and texts) collected from sixty-seven in fourteen groups of third-year students while discussing their Software Engineering (SE) team project at different stages. Then, a natural language processing and machine learning approaches are used to understand factors of abusive behaviour based on conversational mode in a text. Therefore, the following research questions are addressed for collaborative projects in peer-group discussions:

RQ3. Is there a statistically significant difference between online and in-person discussions in terms of polite or abusive language used? Can conversation settings be detected?

RQ4. To what extent can stimulated behaviour shape the understanding and perceptions of peer-group evaluation and consensus in discussions?

To answer the stated research questions, the method is tested at four levels of analysis. First, the research tests behaviour between in-person vs online conversations to find the differences linked to linguistic text features in terms of politeness, abuse, sentiment, text similarity and readability. Thus, the four hypotheses are the following:

Hypothesis 1 *There is a statistical significance between online and in-person peer-group discussions in terms of the **hedging and negation** while discussing the same topic.*

Null Hypothesis 1 *There is no statistical significance between online and in-person peer-group discussions in terms of the **hedging and negation** while discussing the same topic.*

Hypothesis 2 *There is a statistical significance between online and in-person peer-group discussions in terms of the **hate speech and offensive language** while discussing the same topic.*

Null Hypothesis 2 *There is no statistical significance between online and in-person peer-group discussions in terms of the **hate speech and offensive language** while discussing the same topic.*

Hypothesis 3 *There is a statistical significance between online and in-person peer-group discussions in terms of the **text similarity and readability** while discussing the same topic.*

Null Hypothesis 3 *There is no statistical significance between online and in-person peer-group discussions in terms of the **text similarity and readability** while discussing the same topic.*

Hypothesis 4 *There is a statistical significance between online and in-person peer-group discussions in terms of the **emotions** while discussing the same topic.*

Null Hypothesis 4 *There is no statistical significance between online and in-person peer-group discussions in terms of the **emotions** while discussing the same topic.*

Second, the research seeks to investigate how participants evaluate one another based on their perception of acceptability and abusiveness on conversations.

Third, testing online comments from Reddit that were removed is to be sure that the identified textual features that are strongly related to abusive behaviour. In particular, to see whether the politeness, abuse and sentiment features set can be used to predict removed comments to assist moderators before intervention. Finally, advocating that this replication of online vs in-person data should give a broader perspective, particularly, to examine factors of consensus on conversations based on qualitative analysis.

Contribution. The main contributions of this chapter are to:

- (1) Show how characteristics are measurably different between online and in-person group discussion based on the longitudinal observation and textual analysis
- (2) Show why and how users judge contributions differently in two settings

- (3) Perform multiple text classification tasks to identify acceptable and misbehaving features in a discussion
- (4) Investigate key elements of consensus building that affect the nature of the discussion

The outline of this chapter begins by reviewing related work followed by presenting the method and findings for quantitative and qualitative analysis. Finally, concludes by discussing major key points of the findings and limitations of this research.

4.2 Related Work

In this section, the main aspects of this research include (1) factors online disinhibition, (2) peer-group interaction between online and in-person, and (3) capturing online vs in-person behaviours.

4.2.1 Online disinhibition

The online disinhibition represents the circumstances of social constraints and restraints that commonly occur in face-to-face interactions that they do not arise in online settings. Suler [16] proposed six factors of online disinhibition effect, including anonymity, obscurity, asynchronism, dissociative imagination, solipsist introjection and minimisation of authority. There are two primary classifications of behaviour that fall beneath the online disinhibition. These two categories are gracious disinhibition and toxic disinhibition.

Gracious disinhibition defines behaviour in which people might reveal more emotional feelings in online communities than they would in-person, or move out of their way to support someone or offer virtue. Studies have shown that online communication affects the way people behave in-person in conversations in many reasons. For example, Erete [187] found that online discussions about crime impacted the perception of interaction, self-protection and civic engagement that occurred in-person. Similarly, Hendriks et al. [188] investigated the causes of sentiment conversations between online and in-person modes of communication related to familiarity of alcohol drinking

discussion between partners. The findings suggested that familiarity can interplay the discussion and more likely to occur in offline mode of communication. Other examples of examining the harmless and inspiring differences between online and in-person behaviour include political engagement and views [189, 190, 185], civic engagement [186], social support [191], participating in event invitations [185].

Toxic disinhibition, on the other hand, is the behaviour that involves offensive or abusive language, menaces, ministering individuals of pornography, immorality, and brutality in online communities where the person might not do to in reality. Cheng et al. [108] claimed that situational factors before-mentioned as changing mood can encourage or promote people to become trolls. For example, Chandrasekharan et al. [192] addressed the analysis of hate speech in banned online communities to find an adequate method to lessen hate speech based on deviant hate groups. Another study [193] found that motivation effects related to hate expressions and anonymity can help to identify online hate speech. The difference between toxic and gracious is not always obvious in terms of online and in-person communications. However, this chapter aims to investigate the differences between in-person and online conversations amongst peer-groups to uncover significant behavioural patterns in discussions.

4.2.2 Peer-group interaction in online and in-person discussion

The social dynamics of one group can often influence or reflect the way people think and interact towards a particular action or concept. Subgroup in community claim different understanding and perspective of group structure and dynamics, one may interpret college students (*educators*) as a group. Yet, this could also describe a more precise set of students who do not share the same values or beliefs as a subgroup [194]. Cole et al. [182] investigated the differences between online and in-person discussion amongst peer-group students, and found that people are more likely to receive social support online than they would in-person. Also, social support for online and in-person interactions did not lead to an increase of depression or sentiment.

Collaborative projects can face multiple challenges amongst team members including aggression, activity, negation, which would result in poor outcomes and progress. Furthermore, mutual dependence aspects comprising group size or structure can af-

fect group dynamics [195]. Prior work [170] suggested that small group of multidisciplinary and counterbalancing groups in discussion can lead to sampling errors. This works rather, instruct group members to one particular topic that is related to the scope of project so that each can share similar objectives and interests. This is an essential step to limit cultural bias issues. For example, one behaviour can be acceptable in one nation, but it is not in another nation.

4.2.3 Measuring online and in-person behaviour

Amid the recent growth of people on the internet, public discussions are witnessing increased immoral behaviour in both online and in-person communications. Friend and Hamilton [196] presented a study about capturing deception of online and offline communications in a small group setting and reported that online communication enables more trust and sharing personal details than offline communication. To expedite essential conversations on an online platform, most large-scale firms have employed full-time moderators who monitor at least thousands of comments per day [197]. Yet, this is not an optimal approach. One way of overcoming this problem is to study sentiment online behaviours or abusive comments. For example, comments that are offensive, impolite or otherwise prone to make someone leave a discussion. Notably, several studies have developed a wide range of models served anti-social computing research community including features of toxicity [198], conflict and online interactions [106], Bag of Communities [29], antisocial [27] deceptive [47] or disorder behaviours [199]. Yet, a little is known about a direct comparison between in person and online discussions using the available tools for sentiment analysis and other computational linguistics methods. Additionally, they do not permit users to choose which characteristics of abusiveness they are interested in findings, e.g. some platforms may be distinct with abusive behaviour, but not with other forms of abusive behaviour. This can help achieve optimal predictive model to detect such text classification task and select proper features.

This chapter develops an analysis that shows that online discussions are much more prone to precursors to abuse. In addition, the chapter looks at detecting these signals online in a larger data set to see if they are correlated with moderator intervention

through pre-moderated data from Reddit platform.

4.3 Case Study

Prior to this chapter, a case study was presented of similar approach in CHI'19 [170]. The key contribution of the case study work was building a dataset of discussion records for the abusive of behaviour amongst online and in-person conversational data. The ultimate goal was to investigate users' behaviour in small group discussion. The work also examined the impact of a scope of conversation attributes (e.g, text similarity, sentiment, and number of replies.) in affecting view of behaviour in conversations. The preliminary results suggested that the differences in conversation behaviour between in-person and online are measurable.

4.4 Experimental Design

4.4.1 Overview

This section begins by describing the primary method for conducting the longitudinal and textual analysis. The main purpose of this chapter is to look for differences between in-person and online discussions.– particularly, in peer-group discussions. The study targeted a population of third year (of a four year degree) undergraduate students on their professional software development course at the *University of Glasgow*. This population was selected partly for convenience, as the students were already organised into small groups for the purposes of the course and the researcher had access to the students via the course coordinators. However, the demographic of the participants was useful because, as reported in survey results in Chapter 3, most members of online communities users were at age between 18 and 30 which is a typical age range of a college student. In addition, the Pew research center [33] reported that adults who are under 30 years old have mostly faced online harassment.

The main aim is to test for a statistical significance that can reject the primary null hypotheses of *having similar reaction and behaviour has no affect between both settings*.

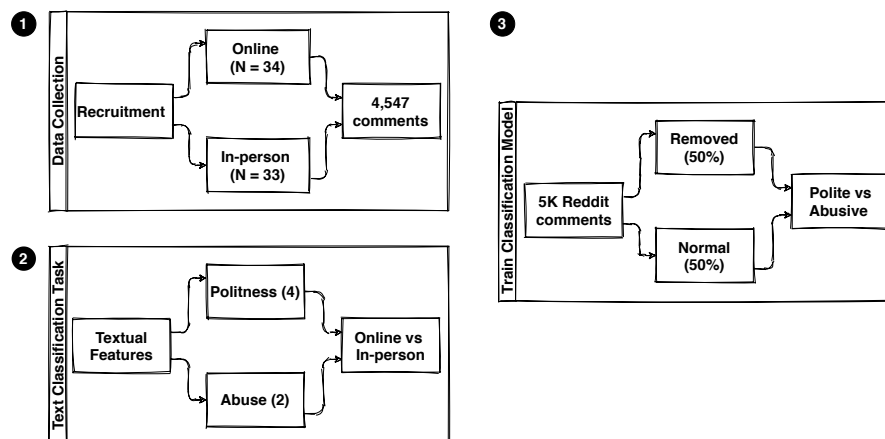


Figure 4.1: Overview of steps for the first part of quantitative analysis.

After the conversational data was collected, a range of textual features were extracted from each comment, indicative of the politeness and sentiment of the comment, as well as readability and similarity to the corpus of conversation comments. A cosine similarity text is used to see whether participants tend to keep focused in the same topic or not online and in-person. A readability test is used to determine the clarity of text. Four features of politeness were used reasoning, reassurance, gratitude and apology. Abusive features were hate speech and offensive language.

A random sample of 5K comments from removeddit platform were collected that showed removed (by moderator), deleted (by user) and visible comments posted on the original Reddit platform. Half of the data contains removed comments and the other half normal (non-removed) comments. The sample of comments was used to show whether classifier is able to predict if a comment is most likely to be removed or not by moderator. The primary purpose of this process was to understand what are the common expectations of acceptable and non-acceptable text in online vs in-person conversation.

An overview of the experimental steps is shown in Figure 4.1. The group size varies from three to five members which each member was allocated to a team by the course organiser. All members have an hour opportunity to talk about their project in each online or in-person meeting. Both conversation settings were audio recorded for in-person discussions and collected from the online group discussion platform. The platform supported all features used to allow effective communications including emoji

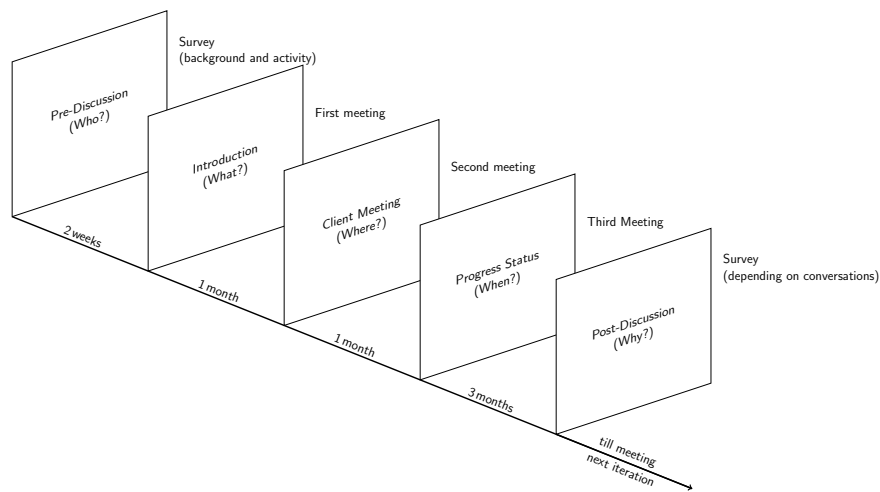


Figure 4.2: Process for conducting the experimental study.

characters, also allow participants to use pseudonym identity if they want. Seven groups were meeting online and seven were meeting in-person. Prior to each meeting, a link was sent to book a time slot to organise timetable meetings for both settings followed by a detailed email about the location and time. In all meetings as summarised in Figure 4.2, participants did not know anything about discussing points in order to allow them freely talk about the progress of project at different stages. Therefore, some points were prepare for each meeting to give them a chance to discuss potential aspects in both positive and negative ways. The topic points were inherited from the course structure and objectives.

4.4.2 Course structure

The primary objective selecting Professional Software Development (PSD) course for the experiment is to link the discussion to the phases of the course. The PSD course is designed to to give students an opportunity to explore real case scenario of software development projects with actual clients. The course is structured to introduce most common problems and techniques at this stage in order to apply them in selected professional projects. The course runs in six phases; each phase is equivalent to one month of different aspects and expectation of the project. The first three phases are

due in the first academic semester and the remaining three phases are in the following semester. In the first month, students are exposed to the offered projects and allocated to a group of three to five members. The second month is about introducing the team members to customer and discuss the project goals. During the third phase, students should be maintaining communication with clients to address related design implication and expectation of the project, and possibly suggest minor features to the customer if needed. The final stage is when the clients are invited to attend a short presentation with a demo of each project. In all phases students are expected to conduct a review progress and meet with a project coach (fourth-year student) to propose or negotiate a project plan, and receive some feedback from their coaches.

4.4.3 Recruitment strategy

The participants were recruited via a public announcement in a school event and email. During the academic year of 2018–19, level-3 software engineering students were invited to participate and 70 out of 242 (29%) students participated in the group discussion study, in which the team project was a part of the degree requirement. The participants were given an information sheet and which describes the research aims about finding the factors of abusive behaviour between online and in-person conversations. In addition, they were given a consent sheet explaining their rights and withdraw from the study if they wish at any time. All participants were informed before participating that they were audio recorded and monitored by the moderator during all live sessions, and the moderator was not in the same room. Participants benefited from taking a part of the study by exploring the projects ahead of their fellow students. The recruitment process took about three weeks before started session assignments.

4.4.4 Pre-discussion (Who?)

Seventy students (96% response rate) completed a pre-survey (see Appendix A.1) for further details about the questions); describing their background and experience of social networking platforms and online activity. The background questions consisting gender, age, education and origin of country. The experience questions cover three

main aspects: activity, familiarity and acceptability on social platforms. Using one unipolar 5-point Likert scale, a variety of questions were formed including daily activity of using social platforms, whether participants have experienced abusive online comments or if they have been forced to delete their own posts and so forth.

4.4.5 Meeting iterations

Introduction (What?). In the first meeting, participants were exposed to all software engineering projects list with descriptive details about each. They had to review all and decide which project to select. There were given set of questions proposed by the course coordinator to discuss with each other. An example of the first group meeting in both settings is shown in Figure 4.3.

Client meeting (Where). This meeting were meant to allow students to discuss plans before after meeting with clients of their projects. The following questions are listed to discuss during this phase:

1. Are we concerned about the scope of the project?
2. Do we have to work across teams?
3. Have we decided to use a new technology (to us)? If so, is it concerning?
4. How are we finding the behaviour of our customers?
5. Are we finding our coaches useful?
6. Have we changed the structure of the team yet? If so, how and why?

Progress status (When?). During the final phase, most students were stressed due to the examination period. So, the goal is to see how they are handling discussion in terms of progress of the project to compare textual analysis with previous meetings.

4.4.6 Post-discussion (Why?)

At the end of the study, participants were asked to complete a short questionnaire contains ten random comments from group conversations and ask questions about

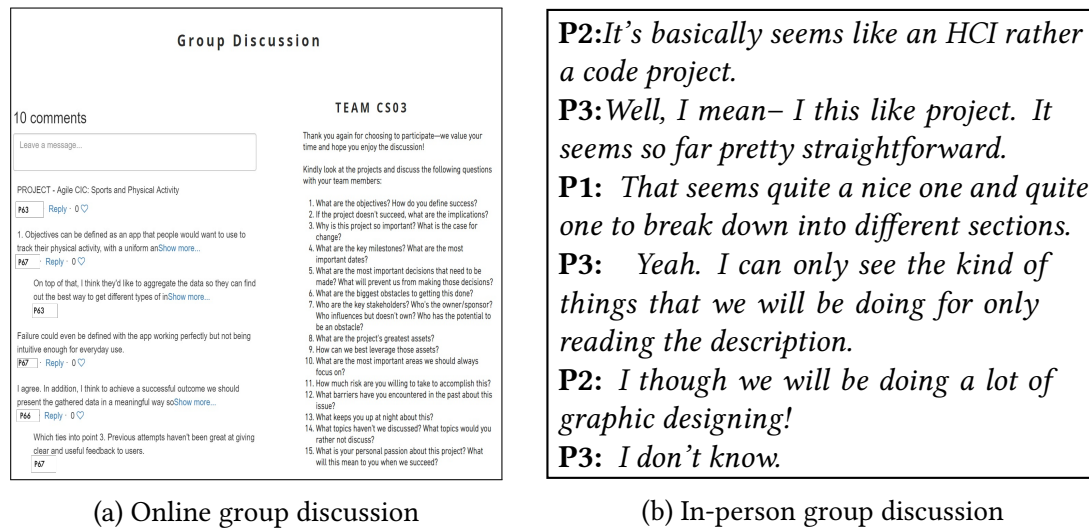


Figure 4.3: Example of online/in-person group discussion form from the first introduction meeting.

each comment (see Appendix A.2 for further details about the instructions). The survey is used to understand the perception and how each participant evaluates abusive behaviour in each meeting individually by asking participants to answer three main questions: (1) What kind of comment is this? (2) Please explain why did you select this. (3) Do you think that this comment was made online or in-person? The disagreement levels were adopted from [200]. The step seeks to find expectations of group discussions, e.g. background or gender perceptions, and how they engage discussions in different group settings, and their understanding on levels of arguments including dis/agreement and abuse.

4.4.7 Ethics

The study treatment obtained ethical approval from University of Glasgow for review purpose (number 300170138). In all cases, conversations were monitored by the researcher. All participants were asked to discuss potentially controversial points related to the project. The information sheet strongly indicated that if at any point in the discussion participants feel uncomfortable they are encouraged to leave or withdraw from the study at their convenience. Also, in extreme situations, the researcher

would end a discussion prematurely if necessary.

4.4.8 Dependent variables

Human factors. As shown in the pre-survey, participants were questioned about their experience with social platforms incorporating abuse factors and measure the independent variable of demographics. To measure experience factors, two sets of questions were used: one set for examining the frequency of daily activity of using online social applications about the likelihood that they meet online friends in-person or witnessing abusive online comments, and the other set to measure the awareness of technology. The frequency variables are activity (Q7), meeting (Q8) and acceptability (Q12). The familiarity variables include membership (Q10), deleted (Q13) and abused (Q14). Each of those variables can provide a pre-measured behavioural analysis of how participants are aware of the abuse-related problems while using social platforms. In particular, it renders a sense of whether background and prior experiences can alter the way they communicate between both settings. It would also reveal group cohesiveness related concerns to understand the attitudes and the nature of similarity and variations among individuals. Since the textual features were numerical, a linear regression test was performed to find the independence between online and in-person behaviour in discussion. In particular, to test null hypothesis for polite, abusive, and emotional textual factors to show if there is statistical significant factor or not between in-person and online conversations.

Linguistic factors. To reveal the behavioural variations between settings on conversations, natural language processing techniques were applied on the analysis to identify textual features demonstrated in 2.3. Specifically, sentiment analysis ¹, hate speech and offensive ² language [25], cosine text similarity, politeness [153], and readability score. Recent studies [108, 190, 170] have shown that those measures are essential in identifying abusive textual activities in discussions. The politeness measures were added to study the stages of how misrepresenting the structure a conversation can lead to provoke incivility and perhaps promoting silence among participants. The package ³

¹<https://textblob.readthedocs.io/en/dev/>

²<https://github.com/t-davidson/hate-speech-and-offensive-language>

³<https://cran.r-project.org/web/packages/politeness/>

Group		In-person (N = 33)	Online (N = 34)
Gender	Male	25 (37.3%)	25 (37.3%)
	Female	9 (13.4%)	8 (12.0%)
Age	18-20	11 (16.4%)	13 (19.4%)
	20-29	22 (32.8%)	21 (31.3%)
Education	High School/GED	25 (37.3%)	22 (32.9%)
	Some College	7 (10.4%)	7 (10.4%)
	Bachelor' Degree	1 (1.5%)	5 (7.5%)

Table 4.1: Summary of participants for each demographic group.

is available in Rstudio that relies on list of 36 politeness features. It can output list of features frequency using multiple metric options: binary, average and count. The function count produces vector of texts that corresponds to each set, then used for analysis of the conversation. The abuse and sentiment APIs produces the output with the estimation accuracy rate (see examples in Table 2.3).

4.5 Results

4.5.1 Participants

A total of 70 participants ⁴ agreed to join all discussion sessions in the study for each iteration. Table 4.1 provides more details on the participants' demographic information. About Seventy-seven percent of participants were male and the rest were female. Participants varied in age from 18 to 39 (66% were aged 21-29). Sixty-seven percent of students were mixed of western and eastern Europeans, the rest were a mix of South Asians. This section relates findings from 67 students who both filled out the survey and attended at least two meetings to a group discussion. The demographic data of the participants is shown in Table A.3.

⁴One group has withdrawn from the study before the first meeting. Their data were excluded from in the analysis

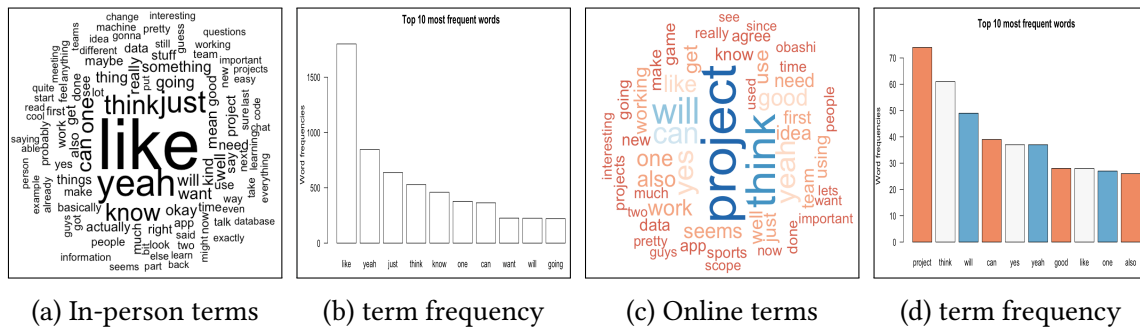


Figure 4.4: Word cloud for online and in-person conversations. All sensitive information, e.g. names or personal information were removed to protect the confidentiality of the participants.

4.5.2 Data set

The collected data⁵ of both settings during all meetings. All group attended the first meeting, only two groups did not attend the online meetings and two on online and in-person discussion during the third meeting. In very rare cases (only once) two members attend one meeting and the rest did not attend that meeting. As shown in Table 4.2, in-person groups were more active and engaging in a discussion (μ 545 comments). In-person groups used less distinct words ranges from (20% - 34%), and online form (46% - 64%). The lexical density was used to determine whether or not the ratio of each comments per group contains more descriptive words about a subject in a sentence, which in this case online groups tend to be more direct in a conversation (μ 60%). Online groups were slightly higher on average using syllable words per meeting. The number of words per sentence were significantly higher online than in-person groups (μ 26%). In-person groups tend to stay longer in terms of time length on average 26 minutes. To explore the most frequent words used in each setting, a word cloud is presented of all conversations with top ten frequent terms as shown in Figure 4.4. Overall, conversations tend to show a degree of agreement. Both settings used the word 'think' most frequently which is part of hedging language. The word was used mostly in online conversations. The word 'can' was also mostly used in online conversations which either refers to questions or for expressing negative feelings with negation. The word 'like' appeared most often in in-person conversations to describe cases or situations.

⁵<https://github.com/aalkhara/Online-vs-In-person>

Team	# comments	# words	# distinct	Lexical %	Syllables μ	Length μ	Time μ (min)
T1	359	2993	890	29.7%	1.46	11.42	16
T2	417	4047	1111	27.5%	1.51	13.16	17
T3	688	5683	1418	25.0%	1.5	10.38	31
T4	464	3855	1064	27.6%	1.52	10.3	25
T5	480	4664	1262	27.1%	1.52	12.08	19
T6	213	1609	556	34.6%	1.53	9.72	22
T7	1192	8329	1694	20.3%	1.47	11.83	41
T8	16	377	249	66.0%	1.65	23.16	12
T9	82	374	243	65.0%	1.68	17.42	55
T10	187	1216	564	46.4%	1.63	21.3	42
T11	15	148	116	78.4%	1.66	22.08	16
T12	145	866	470	54.3%	1.49	34.07	13
T13	204	581	341	58.7%	1.48	15.11	23
T14	87	1103	559	50.7%	1.61	13.43	21

Table 4.2: Descriptive summary of conversations dataset. The first seven teams (T1-T7) were recruited on in-person discussion, and the rest (T8-T14) were online groups.

4.6 Measures of Polite and Abusive Text (Q3)

4.6.1 Experience and background

The frequency questions in Figure 4.5 indicates that 25% of participants reported on the activity and acceptability questions the degree to which they were frequently using social platforms, 72% have met people in-person and 76% have witnessed unacceptable online comments most often. The familiarity questions as presented in Figure 4.6, showed that at least 12% were victim of online abuse, 13% forced to delete own comment, 7% are accepting a friend request with strangers (28% reported maybe) and 63% are active members in focus group on a social platform.

4.6.2 Conversations

The analysis revealed that in-person groups tend to use hedging language (shown in Figure 4.7a) on average more than online groups. T14 was an outlier during the second meeting and were slightly more than the first meeting (an increase of 37.5%), and far decreased during the third meeting by 85.45%. This is largely due to the fact that

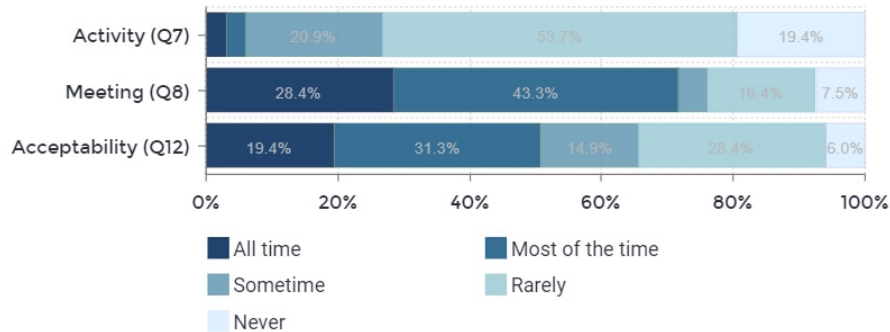


Figure 4.5: Frequency responses of pre-survey for activity, meeting and acceptability questions. The activity question is about daily usage of social platforms, (Q8) indicates the frequency of meeting friends in person and acceptability question enquires the perception of intolerable online comments.

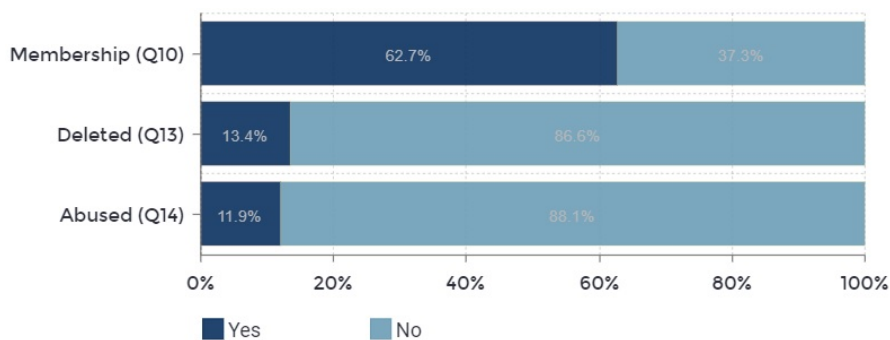


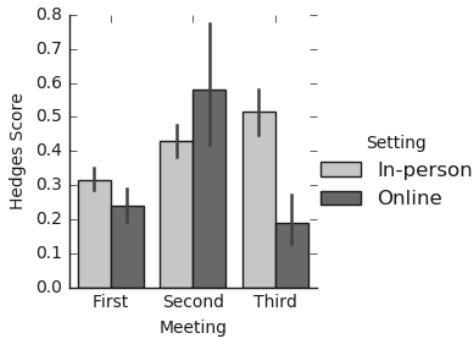
Figure 4.6: Familiarity responses of pre-survey for membership, deleted and abused questions. Membership question asks whether or not the participant is an active member of online focus group, deleted question is know if s/he has been forced to delete their own post and Q14 to see if they have been a victim of online abuse.

this team had equal gender participation during the second meeting. The hedging analysis can confirm that face-to-face interaction increases the likelihood of politeness. The score is calculated by using a vector of text corpus and returns the frequency of politeness features.

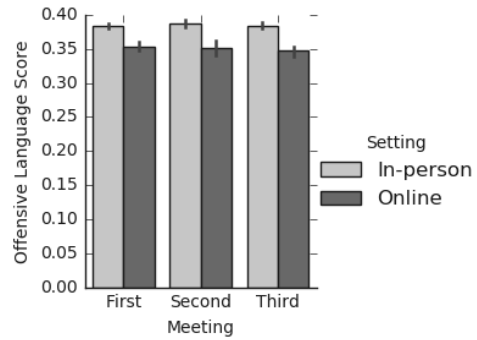
Hate speech and offensive language were statistically significant between in-person and online conversations ($P < 0.05$) as shown in Table 4.3. Figure 4.7b shows the offensive language used mostly by in-person groups. Hate speech on the other hand, was used mostly in online groups. This can validate that possible online features (e.g., anonymity and the quantity of the audience) do not necessarily promote abusive behaviour. Conversely, face-to-face interaction can lessen hate speech, yet encourages offensive language. The score of offensiveness is measuring the confidence of the classifier based on textual features. Online groups were significantly lower than in-person in text similarity ($P < 0.01$) during all meetings as shown in Figure 4.7c. This may indicate that online groups strive to stay on topic.

The readability of text shown in Figure 4.7d signifies that online comments overall are simpler to understand than face-to-face discussions. This is because in-person conversations normally do not use proper grammar or complete sentences. The score indicates the grade level of the New Dale-Chall Formula. Yet, this is an impressive measure of the length of conversations take to reach agreement or consensus. In-person groups in Figure 4.7e showed less positive emotions in all meetings and online groups showed dramatically less during each following meeting— due to the examinations period and approaching the end of the semester.

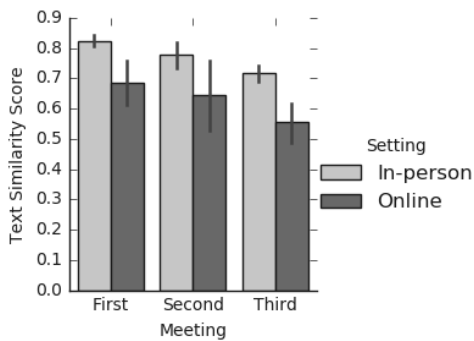
Figure 4.7f display the subjectivity scores of online vs. in-person groups. The subjective text reveals any the confidence score of particular opinions and views of an individual comment. As shown, in-person groups were significantly lower than in online teams. This may suggest that online discussions tend to show greater willingness to express a personal opinion. There was no statistical significance in negation, but hedging language was mostly used by in-person groups. The analysis failed to reject the null hypothesis of negation ($P > 0.1$). A summary of the regression analysis is shown in Table 4.3.



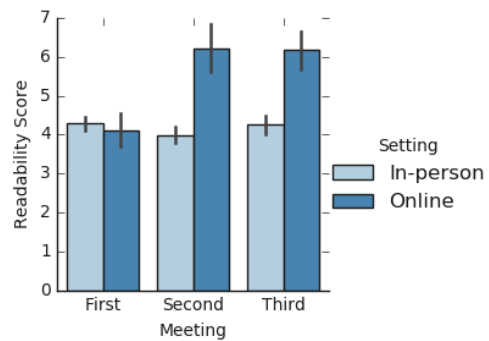
(a) Hedge Language



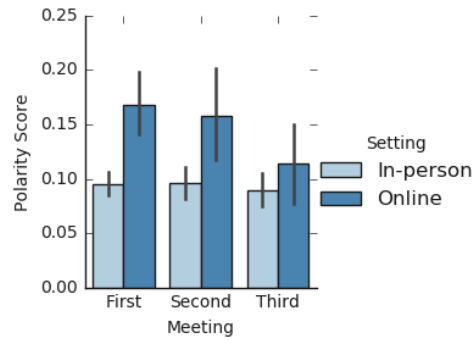
(b) Offensive Language



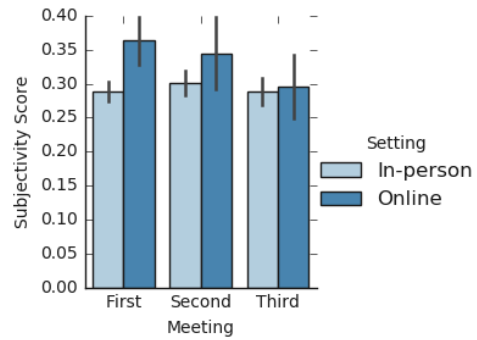
(c) Similarity Text



(d) Readability Text



(e) Sentiment Analysis



(f) Subjectivity Analysis

Figure 4.7: Textual factors among online and in-person discussions during each meeting. Figures (a)-(c) show significant results for in-person groups and (d)-(f) for online groups. Figure (c) cosine similarity in text vectors is used to see which setting tend to stay off-topic. The decimal numbers on the y-axis for the figures represent the percentage of feature frequency. The y-axis values in Fig (d) is the Dale–Chall formula score.

Measure	F-statistic	<i>p</i> -value
Hedging (Politeness)	32.84	< 0.01 ***
Negation (Politeness)	1.58	0.208
Hate Speech (Abuse)	3.87	0.049 *
Offensive Language (Abuse)	45.02	< 0.01 ***
Text Similarity	10.71	0.002 **
Text Readability	7.06	0.007 **
Polarity Analysis	34.85	< 0.01 ***
Subjectivity Analysis	15.91	< 0.01 ***

Signif. codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, $p < 1$

Table 4.3: Summary of statistical significance between online and in-person groups using linear regression analysis.

4.7 Understanding and Identifying Stimulated Behaviour (Q4)

In this section, we investigate how participants evaluate the transcripts from their peers in terms of agreement, disagreement and abusive content using the post-discussion survey. Also, this section investigates the causal relationship between successive comments from textual features to understand the development of abusive content. Finally the section investigates the development of predictive models to identify removed comments based on six selected polite and abusive textual features.

4.7.1 Evaluation of participants

Only 25 participants (37.3%) have completed the post-discussion survey. To begin learning about the stimulated behaviour, a summarised analysis of the feedback responses is reported about their fellow conversions. All participants were given the opportunity to reflect their opinions on ten randomly selected comments from both online and in-person groups, and see how they would evaluate those comments based on three categories of argument as shown in Table 4.4, 4.5 and Table 4.6.

The first comment (C1) is an agreeable statement according to the responses from participants. Almost everyone agrees that the comment is informative and explanatory.









ID	Comment	Category (%)	Justification	Setting (%)
C1	"So it's about external relations which is one of the eight directorates that make up university services and it includes admissions which handle application for all undergraduate and postgraduate teaching courses at the university." (In-person)	▷ Agreeable: 1.00  ▷ Disagreeable: N/A ▷ Abusive : N/A	P34: It provides information which does not seem to be calling for a dispute. P3: Just explaining what is happening. P19: Sounds like someone summarizing what has been said for the sake of clarification. P55 : An elaboration.	Online: (36%) In-person: (64%)
C2	"So my personal passion about this is making something actually working". (Online)	▷ Agreeable: .64  ▷ Disagreeable: .32  ▷ Abusive: .04 	P65: The comment does not contain negative language. P62: Sounds like it isn't going well and that is someone venting their frustrations. P66: Feels passive-aggressive. P64 : Someone is linking personal passion to the successful outcome.	Online: (12%) In-person: (88%)
C3	"I would say we avoid projects which require languages none of us have covered in great detail, such as the C++ requirement for the Medipix project." (Online)	▷ Agreeable: .60  ▷ Disagreeable: .40  ▷ Abusive: N/A	P8: It's an explanation of preferences. P6: It appears to be a counterargument. P9: The comment is reluctant in satisfying requirements. P7 : Suggests not taking a project where the team is not already familiar with a language.	Online: (36%) In-person: (64%)
C4	"I agree. In addition, I think to achieve a successful outcome we should present the gathered data in a meaningful way so that interpretations can be made accurately and efficiently." (Online)	▷ Agreeable: .92  ▷ Disagreeable: .08  ▷ Abusive: N/A	P28: Open to interpretation. P50: Someone making encouraging remarks to expand on an already good point. P49: Succinct point that is just discussing something. P56 : This also makes sense.	Online: (64%) In-person: (36%)

Table 4.4: Post-discussion responses about the first four comments made on both conversational setting










ID	Comment	Category (%)	Justification	Setting (%)
C5	"On top of that, I think they'd like to aggregate the data so they can find out the best way to get different types of individuals to become more physically active ... " (Online)	▷ Agreeable: .92  ▷ Disagreeable: .08  ▷ Abusive: N/A	P10: The speaker is adding an opinion. P11: No dispute. P44: Not disagreeable or abusive. P45: Polite suggestion.	Online: (64%) In-person: (36%)
C6	"Failure could even be defined with the app working perfectly but not being intuitive enough for everyday use." (Online)	▷ Agreeable: .76  ▷ Disagreeable: .24  ▷ Abusive: N/A	28: Sounds correct. P50: Sounds like someone making a differing view about the success criteria for an application. P49: Discusses failure. P56: Usability of an app is an important feature to consider in terms of a products success.	Online: (52%) In-person: (48%)
C7	"I think it's a lot of time, let's find out what all have to say and we'll be done". (In-person)	▷ Agreeable: .40  ▷ Disagreeable: .60  ▷ Abusive: N/A	65: The tone seems domineering but the person is willing to let everyone share their views which is good. P62: Sounds like they don't want to do whatever was suggested due to time constraints or not being worth the time required.. P66: Very dismissive. 64: Approval left to vote.	Online: (28%) In-person: (72%)
C8	"Just like **** said, it's not actually about activate function or input, activate function and input are good, it's about the process, about we actually get there you know following best engineering practices". (In-person)	▷ Agreeable: .64  ▷ Disagreeable: .32  ▷ Abusive: .04 	P16: I don't see a clear argument here. Hence, I have no counterargument. P34: This is something that may lead to a further discussion. P19: Further explaining what someone else said. P55: : An argument about the PSD process trying to make their case for how to get the grades.	Online: (20%) In-person: (80%)

Table 4.5: Post-discussion responses about the second four comments (Cont.) made on both conversational setting.






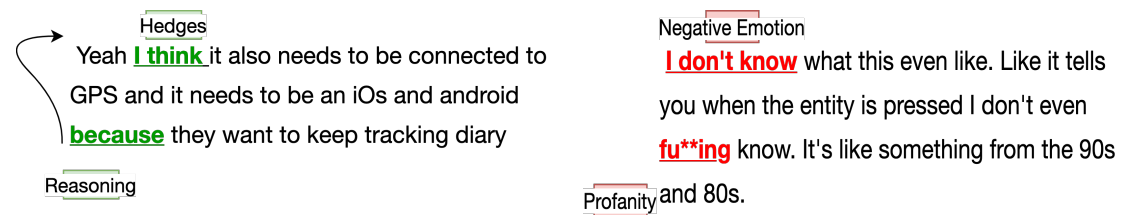
ID	Comment	Category (%)	Justification	Setting (%)
C9	<i>"Oh yeah that's another reason why you don't do Medicare because confidentiality and encryption as much as I want that stuff. This is a little vague".</i> (In-person)	▷ Agreeable: .40  ▷ Disagreeable: .48  ▷ Abusive: .12 	P8: Counterargument. P6: The commenter is agreeing to a view.. P9: Passive aggressiveness. P7 : States the project specifications is vague, which is up to personal interpretation.	Online: (12%) In-person: (88%)
C10	<i>"Well, that just seemed to turn. I think the whole thing is pretty much written as a REST API so potentially they'll just be expanding that maybe?"</i> (In-person)	▷ Agreeable: .68  ▷ Disagreeable: .32  ▷ Abusive: N/A	P28: Opinionated. P50: Contradicting earlier knowledge or understanding, now believe whatever they're working on will only be a single API. P49: Not sure. P56 : Could be either agreeable or disagreeable.	Online: (32%) In-person: (68%)

Table 4.6: Post-discussion responses about the last two comments (Cont.) made on both conversational setting.

Yet, about 36% believe that the comment was said online, and the common aspect of their explanation answer is that they think that online users use less provocative. All participants except three believe that the second comment was discussed on in-person setting, while the comment was originally stated online during the first group meeting. The majority report that C2 is agreeable (60%) and nine people think it is disagreeable. One person (P7) thinks that C2 is abusive stating: *"Passive aggressive tone insinuating that the target of the statement is not working hard enough."* Both P7 and P66 explain the same for this comment, but P66 chose disagreeable rather than abusive. The discussion started by answering about the reason for keeping people up at night about the project. P6 and P9 think that C3 contains disagreeable text, P8 and P7 selected agreeable. About (40%) chose disagreeable for C3. The next two comments C4 and C5 are obvious to most users, particularly C4 since it begins with the phrase "I agree", about nine people think that both comments were revealed in-person.

In C6, about (76%) believe that the comment is agreeable, yet the setting was about half for both online and in-person. For most of those who selected disagreeable, think that this is a counterargument which may lead to further discussion as reported by multiple participants. The majority chose disagreeable for C7, explaining that this is a form of counterargument statement. About (64%) think that the comment is agreeable, one think it is abusive due to the **** notation. This was used to protect the confidentiality of the participants. P16 and P19 think that the comment is disagreeable, yet P34 and P55 believe that the comment is agreeable. C9 was controversial to most of the participants. In particular, (40%) agreeable, (48%) and (12%) abusive. People who selected abusive explain the following: *"Sounds like a cheeky sarcastic remark someone would say talking rudely about someone or down to someone."*(P50). In addition, P32 states *"I am not sure of the meaning but , it should not be directed and be mentioning peoples action directly."* The final comment tagged mostly agreeable for the reason that it is counterpoint to an agreement of argument and suggestion.

C5 was a comment made online, but all participants agreed that this is an *agreeable* comment and said in-person. In C2, we can see that 75% agreed that the comment was made in-person and it is an *agreeable* comment. Yet, P3 reported that the comment was abusive due to the fact that is very sarcastic. One interesting case is that sometimes answers are spilt 50:50, i.e., C3 shows a disagreement in both mode and level.



(a) Hedging feature associated with reasoning (b) Negative emotion associated with profanity

Figure 4.8: Examples of comments associated with polite and abuse textual features. The arrow in each figure indicates the association between textual features in the same comment.

In both cases the justifications which seem to be reasonable is by P1 and P3. Another interesting comment is C8, where comment was said during the third in-person meeting and referred to the instructor suggestions. All agreed that this was a disagreeable comment.

4.7.2 Causality between features

To understand how/why conversations change from polite to abusive language, Bayesian network modelling using the bnlearn⁶ package from R was used to discover the conditional dependency (casual reasoning) between textual features [201]. Each extracted textual feature was represented as a boolean variable to calculate the probability distribution. Specifically, if a comment contained at least one or more instance of a feature, then we treated the feature as present in the comment. Figure 4.8 shows the association between textual features for polite and abusive occurrences in the same comment. In figure 4.8(a), an example of hedging is shown that is followed by an occurrence of reasoning. Similarly, figure 4.8(b) shows an example of negative emotion that leads to reasoning.

Figure 4.9 shows the result of this analysis. The figure shows all investigated textual features using a Bayesian networks graphical model to visualise the conditional dependency between event and evidence variables with associated probability distribution. There is a .51 probability of hedging in a comment leading to the occurrence

⁶<https://cran.r-project.org/web/packages/bnlearn/index.html>

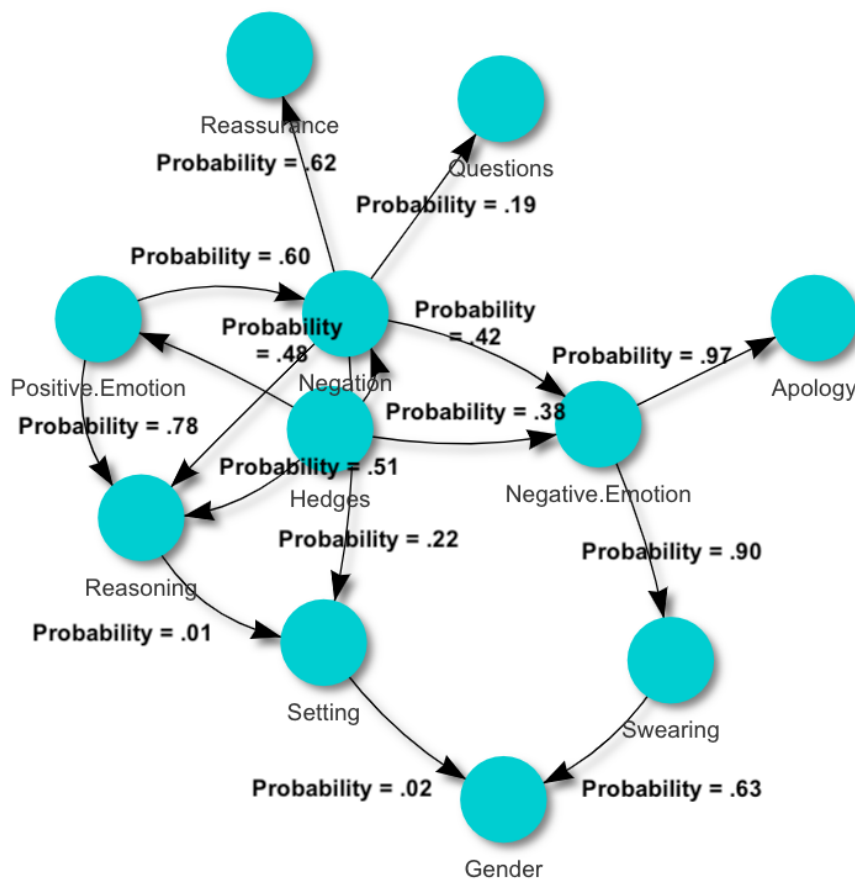


Figure 4.9: Graphical mode of Bayesian networks that shows the conditional probability between each investigated variable. For example, there is high chance of probability (90%) that negative emotion is leading to profanity in the same comment.

of reasoning in a subsequent comment. Positive emotion can lead to reasoning with .78 probability. Reassurance is caused by Negation which has .62 probability of occurrence. There is a high probability .90 to the occurrence of swearing in a subsequent comment.

4.7.3 Classifying abusiveness based on features

With this conversational data, a prediction task was developed where comments were made online (public) and in-person (private). This step is critical to understanding why sort of actions have occurred in online and in-person. In particular, a study has reported that online and in-person have differences and similarities related to hate crime and terrorism that are mostly linked to certain events and radical groups [202]. Another study [203] revealed that people with disabilities were expressing that there is a significant connection between in-person attitude and online behaviour. Both studies suggested that undesirable behaviour can be investigated through content analysis.

Before developing prediction tasks, several textual features were extracted from the comment: profanity, offensive language, reasoning, reassurance, gratitude, apology, TF-IDF and BOW. Those features were selected during the pipeline training step which indicated high accuracy rate for each extracted feature of all investigated textual features. As shown in Figure 4.10, the data set of online and in-person on collected comments is used to find appropriate textual features.

To ensure that classifiers can predict the output of the text before performing predictive model, punctuation, upper case, white spaces, numbers, abbreviations and stop words were eliminated. In addition, lemmatisation and stemming techniques were used to clean text from misspelling and morphological words.

Table 4.7 summarises the results from the classification tasks. The first binary task was to see whether a classifier can be trained to distinguish between online and in-person conversations from characteristics associated with the politeness or abusiveness of the language used. First, Naive Bayes was used to perform binary classification of setting, based on TF-IDF and BOW input features, achieving 72% and 73% accuracy respectively (tasks 1 and 2). Both models did not perform well due to the data imbalance problem. In particular, this may cause one class to dominate another class which is in this case

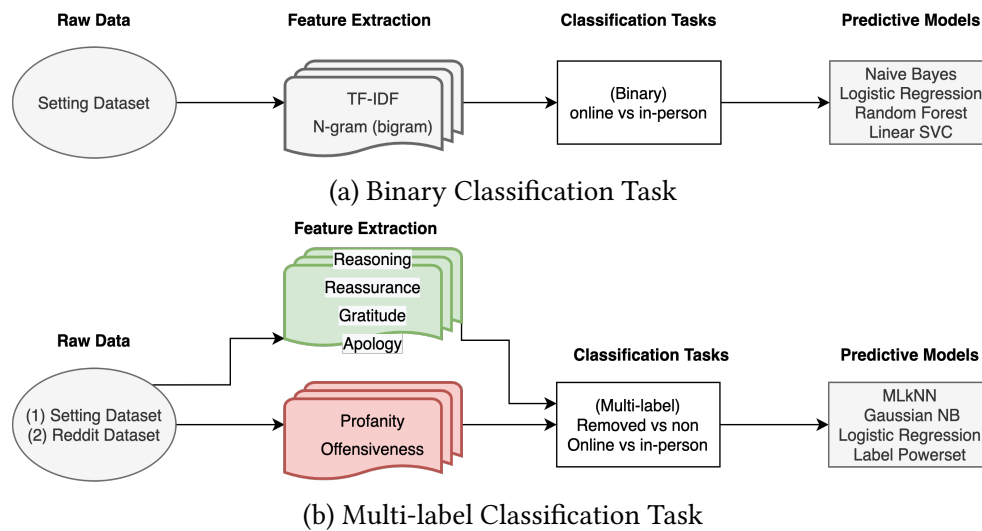


Figure 4.10: A flow diagram of text classification tasks for predicting online vs in-person conversations from our participants and removed vs non-removed comments from Reddit. Figure (a) shows the process of binary text classification and (b) multi-label text classification tasks.

in-person labels. In addition, it can lead to measurement error or sampling bias due to imbalance labels [204].

Next, the following classification tasks were undertaken to detect the setting of conversation in task 3 to task 8 in the table using multi-label predictive models on the selected six textual features: apology, reasoning reassurance, gratitude, offensive language and profanity under ten-fold cross validation. Gaussian Naive Bayes (GNB) and logistic regression were used to transform multi-label independent variables to binary classification problem. Whereas Multi-label k-nearest neighbours (MLKNN) is an adapted algorithm used to predict label set. These classifiers were selected because they were anticipated to perform well for multi-label classification problems. Other approaches, i.e., Decision Tree is useful only with categorical variables and can not perform prediction on collinearity. The dataset does not have normal distribution, so non-parametric models would perform better in this case for a multi-label classification task.

Using profanity and apology to predict setting achieved a accuracy rate of (82% and 86% respectively). The abuse text that includes profanity and offensiveness textual features has accurately predicted whether the comment was made online or in-person using

Task#	Prediction Type (Dataset)	Input Feature Set(s)	Predicted Feature	Classifier	Accuracy
1	Binary (Setting)	TF-IDF	Online vs In-person	Naive Bayes	0.72
2		BOW	Online vs In-person	Naive Bayes	0.73
3	Multi-label (Setting)	(Profanity + Apology)	Online vs In-person	Logistic Regression	0.82
4				Label Powerset	0.86
5		Abuse (Profanity + Offensiveness)	Online vs In-person	Gaussian Naive Bayes	0.83
6				MLkNN	0.84
7	Politeness (Reasoning + Reassurance + Gratitude + Apology)	Online vs In-person	Logistic Regression	0.92	
8			MLkNN	0.93	
9	Multi-label (Reddit)	(Profanity + Apology)	Removed vs non	Gaussian Naive Bayes	0.72
10				Logistic Regression	0.83
11		All (Profanity + Reassurance + Gratitude + Apology)	Removed vs non	Logistic Regression	0.97
12				MLkNN	0.98

Table 4.7: Classifiers performance for predicting in-person/online and removed/non-removed comments using TF-IDF, BOW, abuse and politeness textual features. Classifier that uses abuse and politeness features is able to accurately predict the removed comments.

Gaussian Naive Bayes and MLKNN classifiers achieving (83% and 84% respectively). The politeness feature sets performed well in predicting the setting of conversation based on top predictive features including apology, reasoning reassurance, gratitude. Politeness achieved accuracy of 92% on Logistic regression model and achieved 93% on MLKNN model. This is due to prior distribution which can apply learning relevant information to maximise the approximation of likelihood to find the unseen label from each comment [205, 206] . This means that the computation is more efficient to the memory since it relies on parse matrices.

All the above classification tasks showed high accuracy results in predicting setting of conversation using abusive and polite textual features. Thus, we want to see if a classifier can predict whether a comment is removed or not using the same textual features on Reddit dataset. To begin, we trained removed vs normal comments by moderators for binary classification task on a balanced data set of collected comments from Reddit platform (2,500 removed comments and 2,500 normal comments). Then, the data was used to predict removed comments using Logistic Regression classifier. Overall accuracy achieved was 71%. The total number of predicted removed comments was 3,349; for online setting is 571 (77.8%) comments and for in-person is 2,778 (72.8%) comments. The last four tasks 9-12 show that by combining the top polite and abusive textual features performed the best on all classifiers for predicting removed comments (mean accuracy = 0.98). This is useful to develop online politeness detector by looking for in-person similar to online behaviour, and to see if the feature sets are able to

accurately predict whether a comment is more likely to be removed or not if posted in a discussion platform.

These findings suggest that the abuse and politeness features are valid characteristics to identify the behaviours escalated which may lead to abusive attitude in a group discussion. Nevertheless, the data requires more abusive textual features to capture other undesirable behaviours and become more accurate to detect both conversation setting and removal. So, the analysis suggests that the same features can be used to distinguish between online and in-person settings and abusive and non-abusive content. This suggests that either the online setting is conducive to the development of abusive behaviours, or that the in-person setting is equipped to mitigate their development.

4.8 Qualitative Analysis

Taking "in-person in private" discussions into the "online in public" area can be a compelling method of joining in significant discussions. In particular, to see what features between both setting can interplay the discussion. This is useful to see how a setting of discussion may result in harming rather than polite conversation. In this section, the qualitative linguistics difference between online and in-person conversations are examined aiming to reveal factors of consensus in peer-peer group discussions.

4.8.1 Method

To establish a broader perspective of what was happening during each discussion meeting, a text visualisation tool was used to represent conversations by word tree ⁷ that displays all text of conversations for each setting, then query most common keywords that need to be explored. The plus sign indicates that words that come after original queried key word. For example, the word 'agree' appeared most often in both conversation setting. Yet, each setting has different words coming after the frequent key word; for in-person one of the following came after the key word: because, with or for. For online setting: I think and defiantly appeared after the key word 'Yeah', and 'agree

⁷<https://www.jasondavies.com>

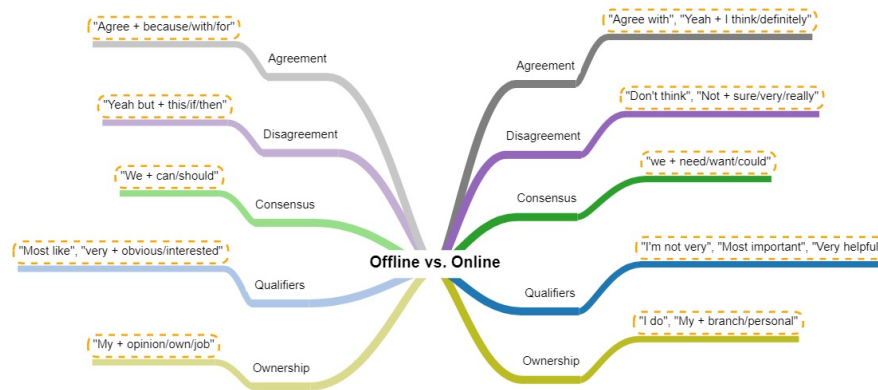


Figure 4.11: Mind map of most cross-frequent phrases related to the addressed themes in online (right branch) and in-person (left branch) group conversations.

with' appeared together. Five major themes occurred more often and led to shift the group discussion and dynamics. These themes include: agreement, consensus building, qualification and ownership. Using some related queries for each theme, particular phrases were observed (sample is displayed in Figure 4.11).

4.8.2 Consensus building factors

(Dis)Agreement. Designing the set of questions for each meeting helped to see how team members can affecting one another while reaching a consensus building. Interestingly enough, some cases of agreement in discussion showed that online/public conversation tend to shut-down the conversation immediately or moves toward another direction as T10 during the first meeting states the following:

(P45): If the project doesn't succeed I think the implications are clear.

(P46): What do you mean?

(P45): That we're either not very motivated or smart

On other hand, face-to-face group discussion tend to show a clear progress of agreement and confirming to ensure that all team members are happy to processed to the following discussion point as what T4 said:

(P15): They want to reach the youth and you know, help them realize [..].

(P14): And it is an app again.

(P15): It is an app again.

Consensus. Online group discussion in team projects can disinhibit social interactions since it is frequently required to be asynchronous while promoting large-scale of participants to speculate their thoughts before sharing or discussing. Yet, in-person discussions tend to rely on physical interactions, i.e., facial expression or body language. That is not necessarily always the case as prior work suggested that consensus building [207] can be achieved by social connections and size of group factors. As an example, T7 starts talking about their positive experience with coaches, then forming a consensus as follows:

(P32): I wouldn't mind actually because I do like systems programming.

(P31): It's not systems programming to do this. It's just basically [..].

(P32): Okay. Alright, that was okay. I guess.

Similarly, online groups sometimes show a process of establishing an agreement. For example, T14 was debating the main concerns and stresses they have about their coach as follows:

(P67): I definitely feel like having 2 coaches is helpful [..].

(P66): I think our coaches have been really helpful so far. I agree [..].

(P66): I think having two coaches is great too. I think they have been [..].

This shows that consensus among groups demands hedging language and subjectivity features to keep the rhythm of a discussion less aggressive regardless of the conversational setting. In some cases, team members will not address hedging appropriately. For instance, T12 discusses the team structure on the second meeting as states this:

(P53): No lol, unless we want to [..] ideas from the other team.

(P57): Yeah I cant imagine that happening.

(P56): Agreed

(P56): OK decided to do it on here instead, scope of the project noted.

Qualifiers. A qualifier refers to a word that emphasises a meaning of a particular word to express feeling, assurance and endurance of conversation [208]. They can also alter the meaning of the context. However, it is meaningful to see whether these were used to express the necessity of confronting ideas or confirming the agreement. During the last meeting, most groups were either satisfied or concerned about their progress on the team project. T3 starts the conversation by asking if there is anything they need to discuss about the progress status. The conversation then diverges as follows:

(P11): I get what you're saying. Like we started off really strongly [..].

(P10): [..] But I don't know if we [..] It would much simpler with django.

(P10): I mean, it's nice. It's a very good technology to know how to use [..].

(P11): A steep learning curve.

T9 mentioned the following:

(P40): I think we will be doing much effort to complete this[..].

(P39): Yes its a bit difficult task but we will try our best to complete this.

In both cases, conversations tend to preserve agreement while disagreeing on a critical point of discussion. However, online conversations lack sufficient argument support or struggle to provide a counterargument.

Ownership. In some cases, one of the team members can blame another member for something did/didn't achieve or not attending regular group meetings, and then imputes everyone else for their weaknesses. It can lead to a destructive consequence and possibly losing trust in the functionality of the team. One team member suspected that her colleague did not use his own script. The other members tried to justify the case to defend their fellow member while he was not present, but still, she was concerned about the situation as stated by T7 during the final meeting:

(P29): What did he send you, sorry?

(P31): he sent me a code that's like a chat bot [..] copied and pasted it.

(P29): It's from the article. I remember I that code.

(P31): Okay.

Online discussions often show that conversations may sway or become less earnest when it comes to ownership. To illustrate further, one member of T13 took a role upon selecting a project on the first meeting, yet the responses were a bit trivial and discouraging. The conversation proceeds as follows:

(P62): I know math is easy [..].

(P58): ☺

(P60): ☺

(P62): I see ☺

4.9 Discussion

A novel empirical approach was undertaken to investigate portions of abuse in group discussions, examining linguistics and human factors. Next, it revealed on the user-level effects of understanding the stimulated behaviour and consensus building. This section discusses the implications of technical design for online discussion followed by limitations of this study and summary of the key findings.

4.9.1 Technical design implications

Dealing with asynchronous-based communication can most often show some lack of activity among users, yet allow multiple people to join a discussion easily. In this study, a pilot version was completed by implementing both synchronous and asynchronous communication methods in the conversation page, (i.e., online group chat and thread-based discussion). Both methods were publicly accessible. It was interesting to see how users were less engaged in the conversation on thread-based version. However, it was a bit less obvious to follow conversations on the chat-based version. For example, it may create significant pressure for some people to respond quickly in synchronous

text-based communication. Yet, asynchronous text-based communication allow people to think ahead and reply to particular comment. In order to circumvent these circumstances, the expectations for participants must be clear in the desired method of communication on the platform. The scale of evaluation categories on the post-survey can also help understand the stimulated behaviour within fellow peers' discussions.

4.9.2 Limitations

Although achieving the stated aims, a number of challenges, concerns, and objections arose in the process of the research. First, it took a significant effort to recruit participants to start and continue in all meetings in both settings. In some cases, some team members did not show up in one meeting at most, which resulted in less time to discuss the assigned materials. The sample is not a representative of all discussion platforms users, yet it provides an insight of exploring the causes of abusive behaviour and validate the approach of collecting, and detecting of conversations based on a particular mode. Additionally, the data is largely imbalance from a number of comment perspective due to the fact that online groups showed less engagement in conversations. One possible suggestion is to provide more interactive web discussion design. Although the initial plan of this study was to run all the six iteration, the remaining meetings were withheld and did not continue holding group sessions due to the incident of cyber-attack towards the end of the first semester, which may cause biasing the sample due to the frustration that student had while completing their development project.

4.10 Summary

In this chapter, a study on how users in group discussion was conducted on team projects to show understanding of group conversations when they need to discuss the stages of their software engineering project. Furthermore, investigated how they behave online and in-person to provide additional context to help identify the changes between different settings. From presenting 67 participants with different representations of a group meeting, the significance of arrangement and discussion were de-

terminated that appeared towards immediately understanding the basis and significance of a group conversation. From these findings, online was different from in-person conversations among groups in many aspects. The qualitative analysis was used to explore the nature of consensus building in group conversations. The classification task was presented to accurately predict whether a text is private vs public or online vs in-person comment.

RQ3. Is there a statistically significant difference between online and in-person discussions in terms of polite or abusive language used? Can conversation settings be detected? discussions?

The answer to RQ3 implies that online and in-person discussions are different in several cases. In particular: in-person groups engaged in a greater degree of consensus building during conversations, through additional hedging of language and being more objective. Conversely, online groups employed more extreme sentiment during conversations and were less concentrated on the prescribed topic. Previous studies have shown that these factors are contributors to additional abusiveness in discussions. it consequently, concludes that the use of online platforms for discussions is a causal factor in the proliferation of abusive behaviour. From these findings, The analysis showed how online differ from in-person conversations among groups, then added qualitative analysis to explore the nature of consensus building in group conversations. Hedging and polarity features were the most significant factors ($p < 0.01$).

Also, the answer provided several classification tasks to accurately predict whether a comment polite or abusive from both settings and would most likely to be removed by a moderator when combining top politeness and abuse textual features.

RQ4. To what extent can stimulated behaviour shape the understanding and perceptions of peer-group evaluation and consensus in discussions?

The answer of RQ4 showed how the stimulated behaviour can vary among participants in terms of evaluating each other due to disagreement. Thus, this will be investigated in the following chapter in details about detecting and analysing polite and abusive disagreement.

Chapter 5

Polite vs Abusive Disagreement: The Case of Polemicists

5.1 Introduction

In the previous Chapter 4, online and in-person differences in conversations were investigated amongst a peer-group discussions. The experiment showed how disagreement evolved differently in conversations between in-person and online settings. The research also showed how the discussion online compared to in-person differed in terms of politeness. However, further investigation is needed to uncover the relationship between polite and abusive disagreement in online settings. In this chapter, a means of classifying disagreement online is developed and evaluated on the top and most active communities on a social networking platform namely Reddit with large scale of data.

Disagreement refers to an argument which individuals construct a different opinion about particular discussion point. It may directly respond to the original argument or become less persuasive. Reasoning and supportive evidences in an argument can impact attitudes of individuals [209]. In some online communities, down-vote might be misinterpreted or misused. For example, a comment on Reddit received a score of 808 points below zero as shown show in Figure 5.1, was reported by the moderator that the the reason for down-vote was mistaken by disagreement [210].

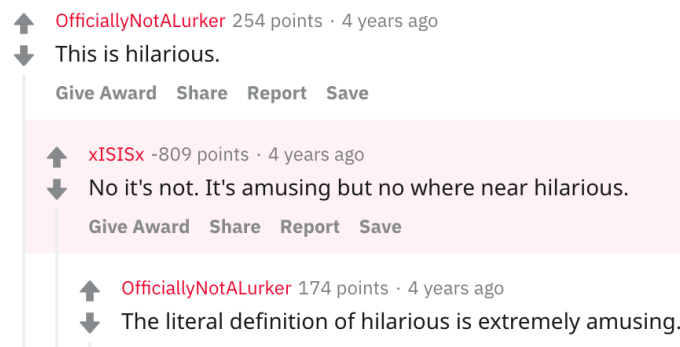


Figure 5.1: An example of disagreement and vote abuse in a discussion dialogue from r/gaming.

This may cause some conflicts on the platform. For instance, many people will down-vote a controversial post because they disagree with it and/or the comments that follow it. Muddiman and Stroud [211] found that rudeness on online communities can prompt abusive behaviour, also profanity is more likely to cause rejection by communities norms. Situational factors were examined in prior research and showed that depressed mood can impact ordinary people to behave just like abusers, and negative behaviour caused by down-votes [108]. Other study [212] examined features of hate speech and objectives in terms of their users activities and concluded that depression and anger are valid measures to identify hate speech . HeartMob [213], a web application developed for people who are particularly impacted by online abuse, and showed that online harassment demands a continuous combination of exposed users' moderation platforms and intervention system design.

People will also down-vote views they disagree with many subreddits will have a sticky post to alert users that down-vote button is not a disagreement button. However, the fundamental problem is that most online communities offer different point systems, and design features that allow community members to provide feedback for posts and comments based on vote scores. Comments that received negative or low votes will lower it on the page and become less evident by default. It appears most often that users are concerned when their posts or comments receive negative feedback without a clear explanation of how and why it happened.

Numerous online discussion or Q/A platforms also operate the same strategy of online voting. The key motivation of this chapter is to *see whether it is possible to discriminate*

between different forms of disagreement, in particular, abusive and non-abusive disagreement. This can help identify the escalation of discussion to design assistive tools and strategies for intervention on moderated-based online communities.

Prior work [214] examined the dynamics of its discussion threads and found that most up-voted comments were posted at the beginning of the discussion. Further work [215] reported that positive or negative feedback from the community is driven by peer-members and it can impact the popularity of a discussion platform. Nevertheless, little is known about the evolution of disagreement and how it affect voting behaviour.

Agreement is an approval position to a continuous conversation reflecting attitude or opinion in a discussion. Disagreement on the other hand, is an oppositional posture [216]. When someone disagrees with another person, this would create a distinct opinion due to the fact that each person has different perspectives, values, and intentions—most of the time disagreement leads to change individual’s view. Teven et al. [217] showed that people who experience tolerance in disagreement are more likely to encounter verbal abuse.

Frances and Matheson [218] argued that disagreement has two types: first one is disagreement by action and second disagreement by the fact of claim. Disagreement by action is when two individuals relate the argument to take particular action, e.g., *‘should we move to this neighbourhood or not.* Disagreement by fact happens when someone is making a claim that relates to a belief, e.g., *I think that the Theory Of Computation course is much harder than the Software Engineering course.*

Another form of disagreement can relate to annotation process amongst crowd workers. For example, in crowd-sourcing applications, Aroyo and Welty [219] introduced a framework Crowd Truth that facilitates improving the quality of inter-annotator agreement and reported that disagreement is inevitable due semantic ambiguity reason. Opinion-based group [220] has shown promising initial results for detecting the disagreement.

In this chapter, a disagreement measure is presented and evaluated to classify abuse and disagreement comments and investigated the factors of vote abuse. Finally, a strategy is proposed to label text for classification purposes. The following research question is addressed:

RQ5. What kind of context enables and promotes polite or abusive disagreement on an online discussion? Do particular kinds of disagreement trigger down voting?

To this end, five main or most active sub-communities were identified in Reddit. Then, five thousand comments were extracted. These comments are labelled by multiple crowd workers to validate and improve the performance for the classification model. Finally, text mining approaches were applied to understand the correlations between abuse and disagreement in a given context. The contribution is as follows:

- Introducing a disagreement scale to detect disagreement in a discussion
- Conducting a longitudinal study analysis on vote abuse on discussion
- Investigating factors of disagreement among contributors on most active online communities

This chapter begins by reviewing the literature, then describing the methodology used for the study and followed by results. The chapters concludes by discussing the implications of this approach and findings.

5.2 Background

Disagreement can be a key factor of online harassment or abuse if unmanaged during a discussion. To better understand the conflicts and consequences of disagreement, the main contribution of this chapter is described, which is the levels of disagreement. The section also provides an outline of the Reddit platform as context for this work.

5.2.1 Disagreement levels

Identifying the differences between disagreeable and abusive comments is a key contribution of this work. Previously, Graham [221] proposed hierarchy of seven disagreement stages to explain how disagreement levels are different. Five from the proposed levels of disagreement were adapted and split the disagreement into two categories

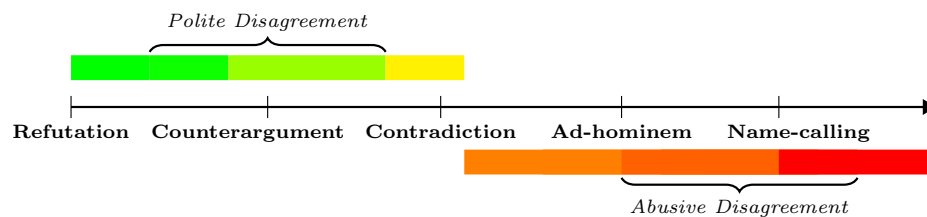


Figure 5.2: The levels of disagreement scale inspired by the hierarchy of disagreement proposed by Graham [221].

for text classification and analysis purposes. as shown in Figure 5.2. The abusive disagreement category spans name-calling and ad hominem attacks. The polite category of disagreement spans refutation, contradiction, counterargument levels. The definition and example of each level can be described as follows:

(L1) Refutation. The author or debater makes a strong argument with sufficient and supportive evidence in refutation level. In particular, illustrating the reasoning behind assertive or counterargument statements. This level mostly involves polite and convincing disagreement.

(L2) Counterargument. In this level, the author has already built contradiction statement, yet with less valid or supportive evidence(s). Counterargument can lead to a weak argument.

(L3) Contradiction. The author is negating the original argument that has no evidence to support the argument. For example, (A): Brexit is the best thing for UK. (B): Brexit is the worst thing for UK.

(L4) Ad Hominem. Rather than responding directly to the argument, the author uses abusive language towards the person who initiates the argument. For example, suppose a governor say: "we need to increase tax rate", another person may reply: "I don't blame you when say something like this because you are the most corrupt leader."

(L5) Name-calling. Abusive language or insults that dose not reflect or contribute to the context directly. This level may include offensive language, hate speech or profanity. For example, "you're stupid!!".

In this work we will use these definitions of different types of disagreement to clas-

sify comments based on text and data analysis. We will then use the classified text comments to understand the relationship between disagreement type and voting behaviour. All the above levels are expected to show escalation in different kinds of contexts in conversations or replies per thread. Five levels into two categories rather than the seven levels were used to make it simpler for crowd workers for reduced reward task to label comments based on the definition and to minimise the risk of data inconsistency as much as possible [222]. Also, the other two levels *refuting the central point* and *responding to tone* are slightly repeated, yet with more complex definitions, i.e., each may fall into more than one category, which can be hard to predict in a context.

5.2.2 Disagreement detection

Misra and Walker [223] examined eight of features of to identify disagreement on 4forums.com discussion platform. The features include politeness and sentiment features such as hedging language and polarity. Hillard et al. [224] presented a classifier that can detect agreement and disagreement and reported that unsupervised learning using n-gram mode with small labelled data can reduce the efforts of data annotation. The authors reported that the topic-independent features show high performance in predicting disagreement of 66% using J48 Trees classifier.

Rosenthal and McKeown [225] investigated the features for capturing agreement and disagreement on for and against side of discussion point in social media dialogue. The detection approach used three-way classification: neither, disagreement, agreement used in a conversation dialogue. The analysis revealed that including lexical and semantic features to identify (dis)agreement achieved, 77.6% on the corpus. Similarly, Yin et al. [226] proposed a three-step method to detect disagreement based on three features: duration, sentiment and emotional. The steps begin by comparing comments to find agreeable and disagreeable comments, then find set of comments by particular discussion topic embodies (dis)agreement collected from the participants. The last step compares both agreeable and disagreeable comments in a broader spectrum of the main topic. The findings suggest that the proposed three features far outperforming the traditional BOW and other NLP approaches.

d'Aquin [227] introduced a preliminary framework to detect dis/agreement in web

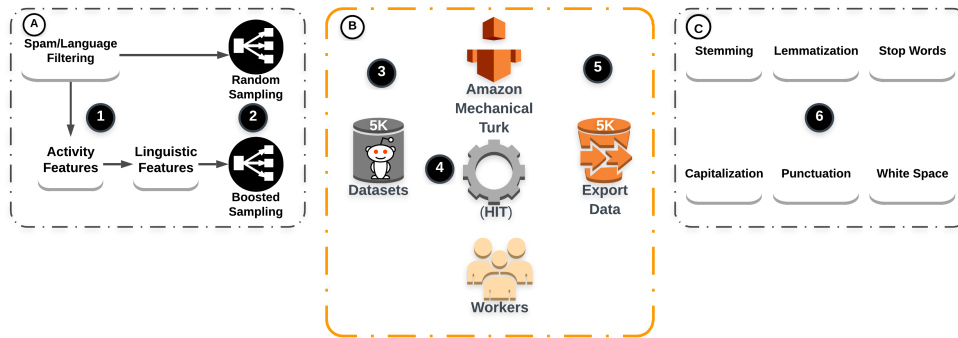


Figure 5.3: Summary of steps for each stage to build disagreement classifiers.

ontologies through measuring controversial and consensus statements. The authors concluded that less consensus implied further mapping between topic domains. Controversial statements, however, require further analysis and mapping related research domains to understand the semantics of disagreement. Prior work [228] used sentiment lexicon model which is based on isotopic to label a sentence in Wikipedia Talk pages and debates. One major lesson learned was that the classifier was unable to accurately predict disagreement when people are using opposite words, sarcasm and contradicting examples. Further research [229] showed in British Exit of European Union (Brexit) that the opinion mining on disagreement or agreement classification outperforms opinion mining used on polarity classification and concluded that Twitter is useful for internet polls.

In contrast, this chapter is leveraging human judgement per comment in order to (1) label the datasets from five disagreement levels as show in Figure 5.2, (2) train the classifier to predict the label of disagreement and (3) classify abuse behaviour based on textual features from the corpus.

5.3 Overview of Methodology

In this section, three main stages are described that are shown in Figure 5.3 for selected data. The first stage (a) data preparation procedure is to provide details on how the data was collected and filtered in order to set up a crowd-sourcing annotation tasks on stage (b). Finally, in stage (c) the preprocessing task is explained for the data before

subreddit	#threads	#comments	Prop (%)	#votes	Prop (%)
r/AskReddit	27,508	41,879	(65.7%)	63,943,538	↓ (83.9%)
r/politics	19,049	48,211	(56.7%)	7,407,555	↑ (12.2%)
r/worldnews	21,473	47,863	(64.8%)	13,217,230	↓ (87.9%)
r/funny	24,686	47,261	(52.2%)	6,057,533	↑ (9.6%)
r/The_Donald	8,944	27,692	(51.4%)	815,343	↓ (82.8%)

Table 5.1: Summary of the collected data from the Reddit API of top one-year submitted posts. The ↓ on the up-votes column refers to voted comments with less than average voting score and ↑ for comments received greater than average score. The gray parentheses indicate the proportional of all submitted comments corresponding to the total of #threads.

the classifier is built *Dataset Description*. The data is collected from pushshift¹ and Python Reddit API Wrapper (PRAW). The PRAW features 11 models; each model is class which contains several attributes to extract data values from Reddit pages. Also, the data was scraped from the removeddit web source to get all submitted pages on Reddit with visible deleted and removed comments. There are more than one million of subreddits and 230 millions of comments stored monthly in pushshift platform. This is a vast number of comments to consider– the study only analyse most active five subreddits ranked by the highest total number of subscribers of each subreddit. The collected data shown in Table 5.1 summarises the selected subreddits that contains all users’ activity from top, controversial and hot posts from 2017-2018. There are more than 213K comments of 125K unique conversation threads from top most active five subreddits on reddit. The average posted comment per thread is two comments.

5.3.1 Preparation procedure (A)

As in Figure 5.3, the methodology is elaborated by steps for each stage. In first stage shows the preparing the data in step ❶ before sampling in step ❷. In particular, extracting three pages or more of top, controversial, hot and rising from each subreddit. Further details is described as follows:

Spam/Language Detection. Comments in step ❶ that show some web links, incomplete or non English sentences and advertising or malicious activity are removed before

¹<https://pushshift.io>

sampling the data.

No. of Words. Comments that contains less than two words were removed from the samples to gain more meaningful and complete context. Also, this will allow the classify to learn and accurately predict the label on any text classification model.

Remove duplicates. Any comments with similar text were excluded to enhance the variety of words list and learning process during classification tasks.

Activity Features. Proportion of deleted/removed and vs. non-deleted posts and sores which is accumulated by subtracting down-votes from up-votes [230]. Non-deleted comments include even where the number of comments per thread is even and odd where the number of comments per thread is odd. Removed comments are done by moderators and deleted comments by users.

Linguistics Features. Various factors of text analysis were used that incorporate abuse, sentiment and politeness features. The politeness classifier [153] uses 36 politeness feature sets ². The package is available in Rstudio and compatible with the python library SpaCy. When it runs, it returns a count of each feature. The most relevant features to this research that are used are listed with examples in Table 2.3.

The abuse classifier ³ include sentiment profanity (swear words) [231], hate speech and offensive detection [25]. For sentiment analysis, TextBlob ⁴ was used to get positive, negative and subjective feelings. Politeness features are hopeful as well to show the spectrum in conversations in terms of efficient communication and healthier argument. The politeness features include reasoning, negation, gratitude and hedging. Another measure is Dale-Chall readability score [232], this can show the complexity metric of how difficult the sentences in each comment.

Sampling. In step ②, a total of 5000 comments were selected(2500 boosted and 2500 random sample). Activity and linguistics features were applied to include highly abusive and argumentative comments from most active five subreddits. Theses samples are essential to prepare for the crowdsourcing step to perform the classification task. The collected samples shown in Figure 5.4 describe data examples of polite, abusive and normal by subreddit and status of comment. Boosted sample has larger score

²<https://cran.r-project.org/web/packages/politeness/>

³<https://github.com/t-davidson/hate-speech-and-offensive-language>

⁴<https://textblob.readthedocs.io/en/dev/>

of profanity/swearing than random sample, mostly in `r/the_Donald` followed by `r/politics`. Deleted comments are slightly higher than normal comments in swearing plot. Negation is mostly used in `r/AskReddit` followed by `r/politics` then `r/worldnews`. Most negative comments in the boosted sample tend to be routinely removed by moderators. Finally, hedging for boosted sample is used most often in `AskReddit` and `worldnews` subreddits on viable comments. Also, boosted sample contains higher number of down-vote on average.

5.3.2 Crowdsourcing set-up ③

Using dictionary-based words only in any approach can lead to bias problems in data analysis due to the difficulty of misunderstanding non-dictionary-based words, e.g. 'yak shaving', which refers to useless task that cause people to do recursive tasks. This occurs when a classifier is trained on most frequent deleted comments that caused by implicit biases and may not necessarily sway threats from abusers over time while interacting [72]. Thus, the same sampling techniques were followed [27, 26] by randomly selecting 1000 comments (500 between deleted and removed, and 500 normal posts) of each subreddit; 5k comments in total for step ③. Each sample contains comments of high and low scores of votes; from each sample top, controversial, hot, and raising were extracted. In step ④, describes a crowd-sourcing experiment that asked workers to label the sampled comments based on the the five disagreement levels [221] (name-calling, ad hominem, contradiction, counterargument, refutation) on Amazon mechanical Turk (as example in Figure 5.5. Each comment is at least rated by three independent workers and averaged by applying a statistical measure krippendorff's [233] alpha to estimate the inter-rater reliability. The strategy was constructed in process as shown in Figure 5.6.

Pre-labelling. Prior to publishing the batches on Amazon platform, a pilot experiment was conducted by using a 10% of the sample samples (500 comments) from the 5K data and asking two groups to do the annotation task based on the disagreement scale. Each group has three workers who labelled 250 comments that were randomly selected from both random and boosted samples. The purpose of this step is to ensure that the instructions are clear and that the annotated data represents a significant ratio of the

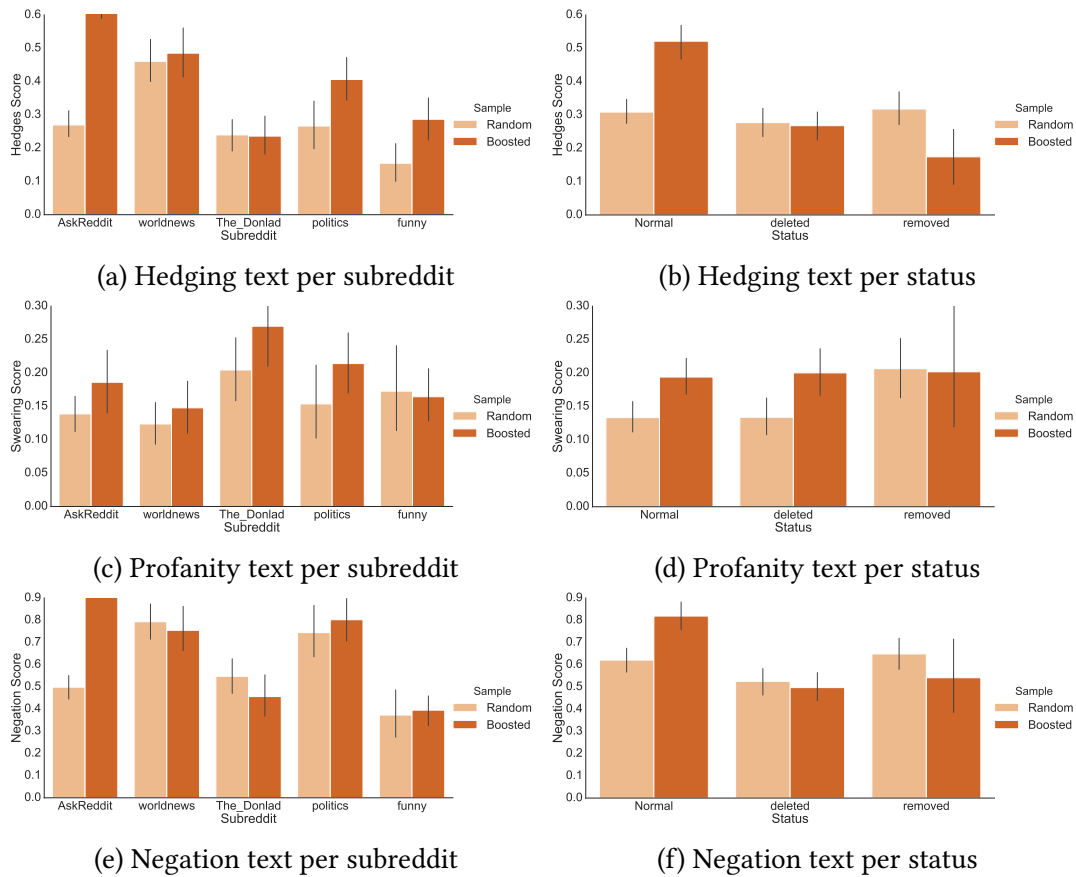


Figure 5.4: The percentage of samples with selected features for random ■ and boosted ■ Reddit comment data sets. Left column of Figures (a),(c) and (e) describe samples of top three significant textual features for selected subreddits, and the right column of figures (b), (d) and (f) show the same textual features for the status of samples. The y-axis represents the percentage frequency of each feature in the given sample. The upper bound with high value indicates higher score of feature frequency and lower bound indicates smaller score of feature frequency per subreddit or comment status.

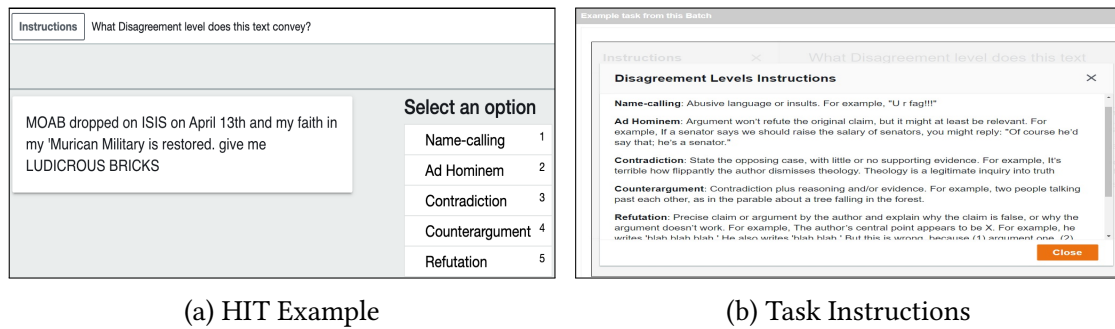


Figure 5.5: An example of HIT task with instructions of how crowd workers were able to label each comment from the dataset.

samples.

Assignment. The number of HITs per assignment can play a significant impact in quality of answers. Prior work shows that a task demands numerous efforts for high standard results can be accomplished by measuring the impact of time on the number of tasks [234]. Most recently, CrowdEval is a combined method of gold evaluation and peer evaluation was proposed to measure workers' performance and reliability [235]. These approaches of errors and reliability led this research to divide tasks into three batches shown in Table 5.2.

Performance. Krippendorff's alpha [233] and Cohen's kappa [236] are effective statistical measures of agreement among raters or judges. The difference between the two measures is that Cohen's kappa is used for less than three raters and Krippendorff's alpha is used for two or more raters. In our case, we have three raters assigned per annotation task. Using at least one Inter-rater reliability measure is essential to measure the agreement amongst annotators that produces a value ranges from 0 perfect disagreement to 1 perfect agreement. The alpha measures reliability coefficient of agreement, and also is associated with content categorisation. The test results for reliability of answers on all submitted batches achieved on average $\alpha = 0.78$ using krippendorff's measure. If the α is less than 0.67, then this indicates a low inter-rater reliability due to statistical significance which the confidence interval is rejected below the probability of selecting distinct answers [237]. The average time for labelling a comment is one minute and 80 seconds. A 12-hour time limit was assigned and most cases the deadline was extend for another 12 hours.

Round	#batch	#comment	#HIT	#worker
1st	11	50	150	129
2nd	10	50	150	136
3rd	8	500	1500	400

Table 5.2: Task assignment of each round per batch. HIT is a comment that is labelled by three distinct workers. The reward value is one cent per HIT/comment

Results. Upon submission of each batch and answered at least by three independent workers, the submitted results⁵ were reviewed and approved the assignments. A qualification requirement was added in the next batch after approving the previous batch to filter the list of workers who have completed the task before. This is useful in most cases to ensure that the data receives diverse opinion from multiple workers and to limit biasing the sample as much as possible. If the results were not effective in terms of the agreement, the batch was resubmitted and extend the deadline for another 12-24 hours. In the second batch, 126 workers were rejected since they have participated already or received lower score in the inter-rater reliability test and thus did not met the qualification requirement to complete the task. The final batch had 125 workers that were rejected due to the not meeting one or both of the qualification requirements. Finally, the result in step ⑤ is used to train the classification models to detect disagreeable and abusive comments.

All the strategies described above including inter-rater reliability measure Krippendorff's alpha and distinct number of workers for performance of task known as qualification test are crucial gold standard elements to ensure that the Turks were engaged in the task and provided an acceptable answer.

5.3.3 Classification ③

To avoid the problem of inconsistency during the classification phase, six significant preprocessing techniques were used including stemming, lemmatisation, stop words, all caps, punctuation and white space removals during the last step ⑥ before performing the classification tasks. Since most comments posted on Reddit have used improper language and grammar, preprocessing step is so critical to maintain generalisation and

⁵<https://zenodo.org/record/4632805#.YFrxEZMzbyg>

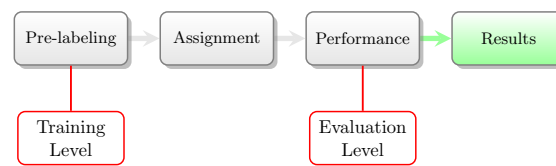


Figure 5.6: Precautionary steps in crowdsourcing experiment.

limit over-fitting the classification model. This will ensure that the text in a given comment is predictable. Three main classification tasks were leveraged: (1) binary (2) multi-class (3) multi-label classification. Firstly, predictive models were trained from the collected labelled data to first predict if a comment contains abusive disagreement (L4-L5) or polite disagreement (L1-L3) using Naive Bayes classifier. Secondly, multi-class algorithm (random forest) was used to only train the disagreeable levels (L2-L4) on the comments. Finally, comments were extracted by abusive and polite features along with all levels of disagreement using multi-label approach, e.g, Multi-Label k-Nearest Neighbour (MLkNN) or Binary Relevance.

5.3.4 Ethics

Since the analysis contains some deleted and removed comments, ethical approval (Application No: 300180163) was obtained to carry out this research and to reduce potential risks of the collected deleted data. In all cases, the user's identity and other sensitive information were removed to endure the anonymity of redditors.

5.4 Understanding Factors of Disagreement and Identifying Abuse

All labelling tasks were completed for the 5K comments on each batch. The comments were mostly contradiction followed by ad hominem and counterargument. Name calling and refutation were labelled much less. The most occurring labels were contradiction (53%) and Ad Hominem (25%) followed by counterargument (19%) and name-calling (2%). The refutation level comments were only representing 1% of the labelled

samples. Most of the comments (61.7%) were labelled similarly of at least two workers agreed on the disagreement label of the comment, out of which (10%) all the three workers agreed on selecting the the same label. In this section, the result of the analysis is reported to reveal factors of disagreement and understand how abuse in online discussion is developed. Finally, the performance of the classification tasks is shown to identify disagreeable text based on abuse, politeness and other textual features.

5.4.1 Factors of disagreement

Context. After the labelled comments was collated from the crowd workers, a text analysis was performed using textual features of abuse, sentiment, and politeness described in Section 5.3.1 during the data preparation on the first sage. The comments of L1 and L5 were excluded since they were labelled significantly less than the other three levels. Also, both levels were either too polite or too abusive. Thus, the analysis is focusing on L2-L4 to uncover the escalation of conversation and see how and when it starts politely then reaches the limit of abusive comments. The disagreeable three measures are: ad hominem (L4), contradiction (L3) and counterargument (L2) as shown in Figure 5.7 considered all politeness textual features, yet reported the most significant results including reasoning, gratitude, hedging and negation textual features.

Using hedging in Figure 5.7a and gratitude in Figure 5.7b features in a context were more likely to appear on the first two levels of disagreement. This shows how conversations tend to shift towards abusive language when using less hedging and gratitude features. The disagreement scale contains several aspects of negations. Therefore the negation language measures a content to see that point A contradicts point NOT-A. The Figure 5.7c indicates that the negation score were used much more in the counterargument followed by contradiction levels of disagreement. This can validate that content that contains more negation features is more likely to reach persuasive and polite disagreement. Each disagreeable level compose stages of supporting an augment. To understand how the disagreement provides enough supporting sentences, the the reasoning measure was used. The Figure 5.7d shows that the first level provides enough evidence to support the disagreement in an argument.

Abusive textual features reported in 5.8, that is hate speech as shown in Figure 5.8b,

profanity (swearing) in Figure 5.8c and offensive language in Figure 5.8a were used much less while confronting central points of discussion on the second and third disagreeable levels. It was expected to see higher scores of hate speech in ad hominem level since this contains abusive or vulgar language. On hourly basis everyday, it seems that on average hate speech, offensive or profanity languages occur most often during the early hours of morning when possibly moderators are less active in the community. This suggests that people who would confront or discuss ideas would less likely use profanity terms while arguing. Surprisingly, comments that were classified as counterargument show higher scores of abusiveness and carry higher scores of negative emotion in Figure 5.8d. This means that comments that contain negative emotion over time are more likely to become abusive

In terms of sentiment analysis as shown in 5.9, positive and negative feelings in Figure 5.8d and 5.9b appeared more in the second and third disagreeable levels. Meanwhile, the first level tend to include wide spread of perceptions in the content. All scores for Figures 5.7, 5.8, 5.9 indicate the confidence of classifying each feature. The subjectivity in Figure 5.9a is almost higher in terms of score in the ad hominem level. This may suggest that users tend to be more subjective yet abusive as suggested by recent researches [238, 239]. The Figure 5.9c shows that when elaborating using more words as shown in Figure 5.9d, the readability of the text becomes easier to read. A linear regression analysis was used to show the dependence between disagreeable and the textual predictor variables. Multiple factors of text analysis were performed that describe abuse and it shows that the hate speech is statistically significant ($p < 0.001$) and offensive language ($p = 0.002$). Hedges, negation and subjectivity ($p < 0.001$). The results suggest those textual features are predictable measures for identifying disagreeable comments.

Duration. Another important factor is response time per unique conversation or discussion thread. Overall, users are more active during the night hours and comments posted during morning time (Eastern Time Zone) take less time to reply to, particularly in L4 (mostly afternoon) and a longer time during the morning. It takes users more time to reply during the night peak hours when using polite disagreement (L2-L3), Yet, it takes much less time to reply when using abusive disagreement (L4). The conversation can easily escalate from L2 to L4 based on the duration to reply, and begin to decline (cool down) during the morning with less time to reply in L2 and L3.

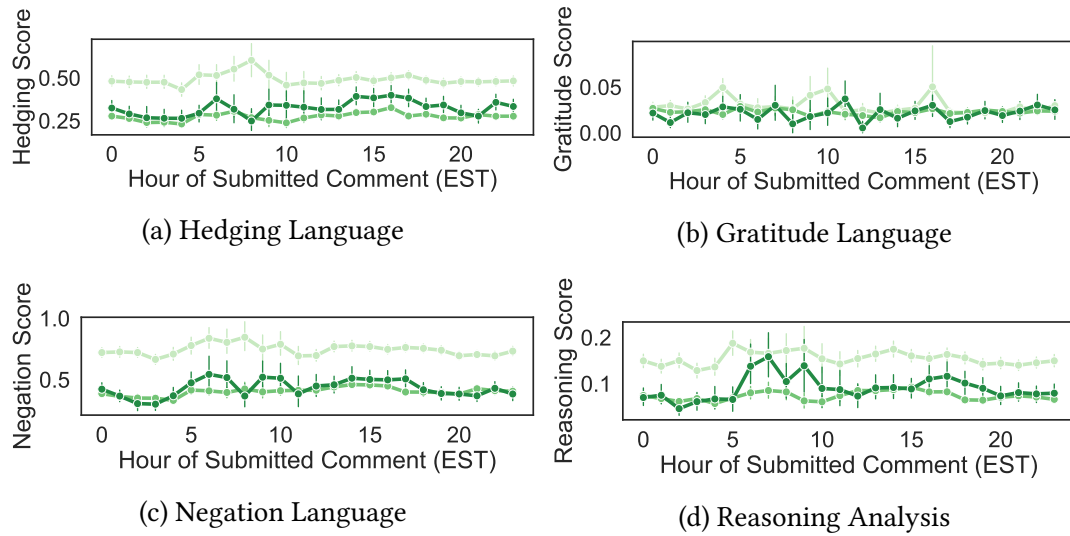


Figure 5.7: Hourly submitted comments based on the four politeness features by disagreement level L4, L3 and L2. The features are hedging, gratitude, negation and reasoning. The y-axis shows the confidence score of factor.

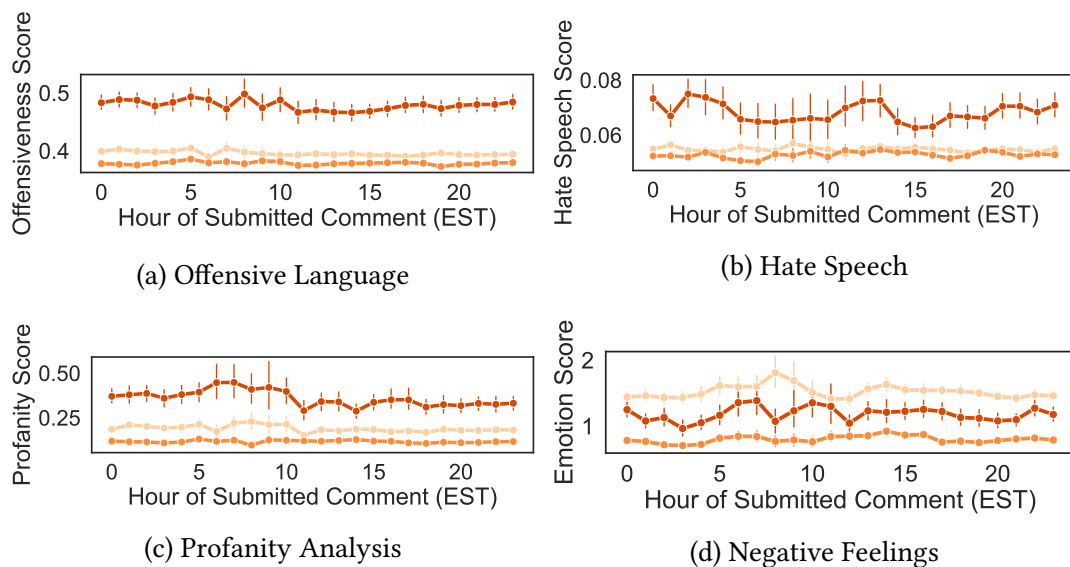


Figure 5.8: Disagreement by abuse textual features for levels L4, L3 and L2. Abusive features include hate speech and offensive language, profanity and negative feeling. The y-axis shows the confidence score of factor.

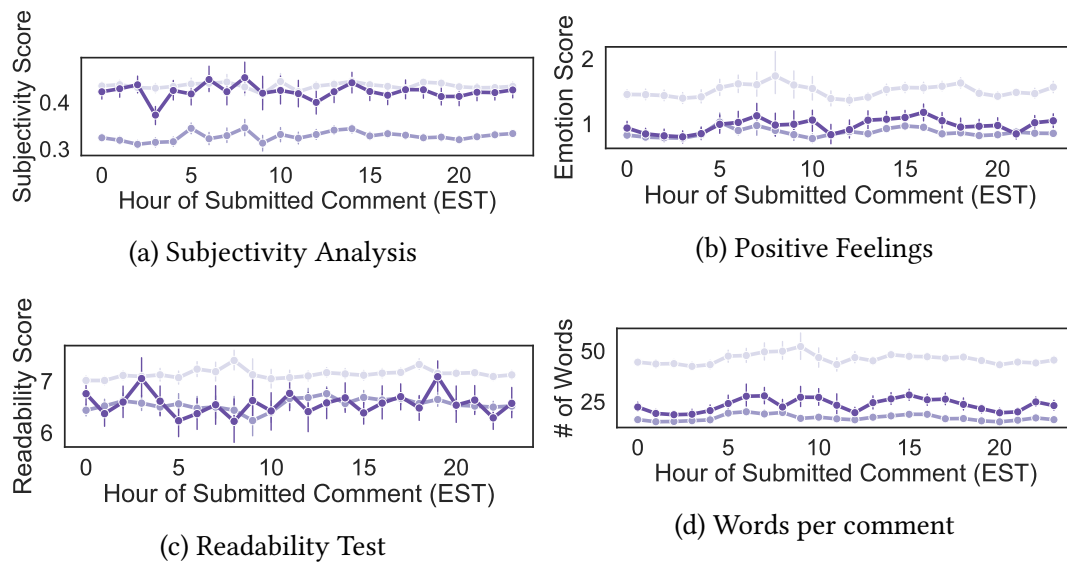
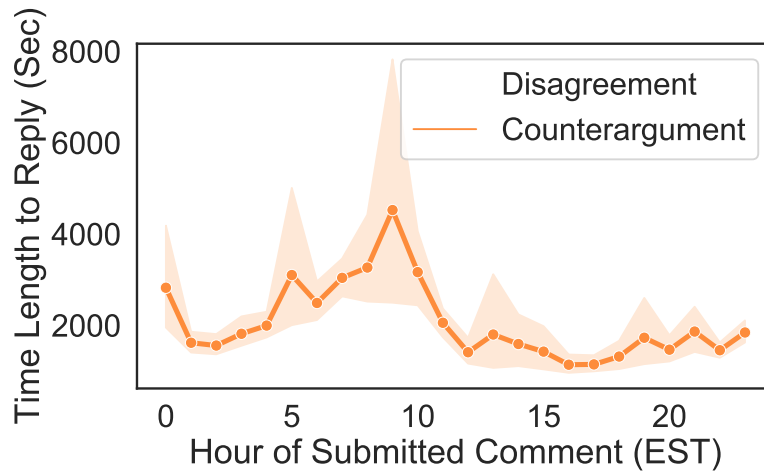


Figure 5.9: Disagreement by sentiment with textual analysis for levels L4, L3 and L2. The readability test scores (k) uses the Dale-Chall readability formula. The y-axis shows the confidence score of factor.

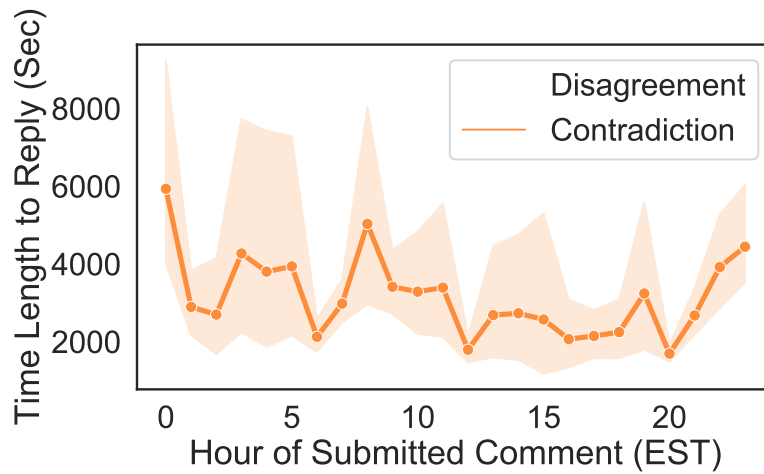
On a daily basis, abusive comments tend to appear more quickly during weekdays and emerge more slowly at weekends. It takes less time to reply with polite disagreement (L2-L3) during the weekend. This may suggest that while the censorship in online content is less active during the and normal hours on weekends, abusive comments tend to appear more quickly during week nights.

To gain better perspective on this and to see how long conversation can last when being abusive, each disagreement level was examined against the duration of time. As shown in Figure 5.10, when the respond takes longer time on counterargument and contradiction levels, comments that are on the ad hominem level become quicker to respond. In addition, ad hominem comments tend to be actively posted as quickly as possibly during the morning time.

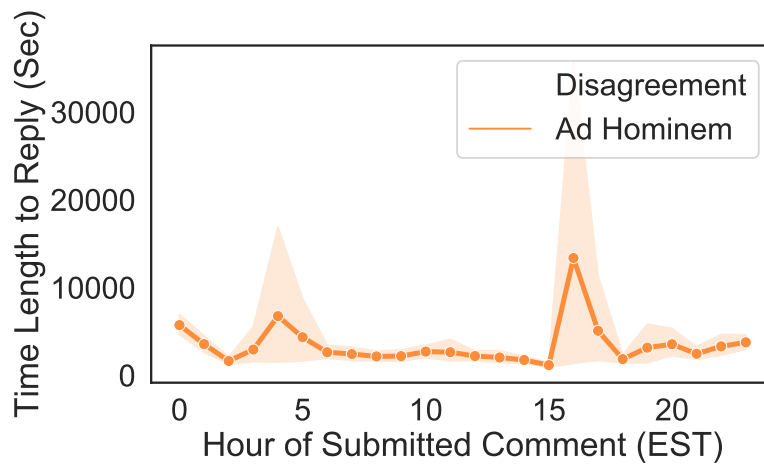
This phenomena has been described by Godwin's law which suggested that if an online discussion takes longer period of time, someone will eventually attempt to compare another person or something to Adolf Hitler to end the discussion[240]. This is an obvious form of ad hominem behaviour when people attack the person who made the argument rather than responding to the augment itself.



(a) Counterargument Level



(b) Contradiction Level



(c) Ad Hominem Level

Figure 5.10: Response time of disagreement levels

Status. There are three different types of status of each comment: deleted, removed and visible. Removing comments can only be action by moderators, whereas users can delete comments. Visible comments are not deleted nor removed. Most communities have different rules. In Table 5.3, an example that provides comment of each disagreement level is shown.

The most deleted comments were found in (L3) contradiction (50%) and (L4) ad hominem (25%). The removed comments were identified mostly in the contradiction level (53%) and ad hominem (26%). This may suggest that users' comments might be removed by moderators or users delete their own comments due to disagreement.

5.4.2 Vote abuse

Most comments that were labelled disagreeable were on the first three levels. However, some disagreeable comments were (-/+) voted not necessarily due to the contribution of the discussion as shown the examples in Table 5.3. In particular, 196 deleted comments received (-) votes below 0, mostly in contradiction level (59%). The removed comments which received (-) votes were 151 and occurred mostly in contradiction level (47%). The average score in the sample data is 170 of which 12 deleted and 2 removed comments. Removed comments were in the counterargument, and deleted comments (50%) in contradiction.

This confirms that the purposed disagreement levels model is valid to distinguish between abusive and non-abusive content. If someone disagreed with the views or opinions are formulated in a comment, should the relevant feedback be to (-/+) vote it, consequently reducing its visibility in the discussion? Is that because they liked or disliked a comment? To answer this, a small test was performed by checking all down/up-voted comments in the sample and found that many disagreeable comments were down-voted due to disagreement, and some abusive comments received higher score of up-votes. Most abusive comments received higher up-votes during early morning time and afternoon times.

Usually there is a clear pattern of up-votes, i.e, L4 level of disagreement comments that are highly up-voted show a clear decremental number of up-votes in L3 and L2 of disagreement over time from the community. So, this implies that the number of

Level	Comment	Subreddit	Status	Score
Refutation ✓	Maybe nothing happens because if we don't see the demonstrators walking past window we go back to binging Daredevil. Hopefully this will convince enough people they need to get out and demonstrate even if its on their own, because then the rest will follow.	r/worldnews	Deleted	-2
Counterargument ✓	Lol. "I know I keep being wrong, but it doesn't matter anyway because of another thing I am wrong about".	r/worldnews	Removed	0
Contradiction ✓	No. This is why America is going to fall.	r/The_Donlad	Deleted	1
Ad Hominem ✗	This is some Nazi level shit everyone involved in this would IED'ed or shot to the head. He said without the slightest hint of irony.	r/AskReddit	Visible	15
Name-calling ✗	That's stupid	r/funny	Deleted	1

Table 5.3: Example comments with details of each disagreement level labelled by the crowd workers. The first three levels are disagreeable and the the last two abusive comments.

up-votes is often misused with disagreement. In addition, the number of words used in comment is much significantly higher in L2 followed by L3 then L4 during the weekend (SAT and SUN). This may suggest that the L2 comments showed that users tend to provide more explanation when supporting an argument at the peak of active editing and moderating days, while redditors who submitted comments that fall under L3 and L4 use less words.

5.4.3 Capturing disagreement

Using supervised learning algorithms, binary, multi-label and multi-class classification methods were used to detect disagreement. As shown in Figure 5.11, features extracted were sentiments associated with abusiveness and politeness, Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BOW) from each comment in the data. Classification was used to predict polite and abusive disagreement according to the levels that were labelled by the crowd workers.

First, the data was prepared and cleaned as stated in preparation step in section 5.3.1, then extracted textual features using TF-IDF approach to implement binary classification task using a Naive Bayes classifier under 10-fold cross validation to predict if the

Prediction Type	Input Feature(s)	Predicted Features	Classifier	Performance
Binary	TF-IDF	L1-L3 vs L4-L5	Naive Bayes	62%
Multi-class	BOW	L1-L5	Random Forest	53%
	BOW	L1-L3	Random Forest	73%
Multi-label	Sentiment (Polarity + Subjectivity)	L1-L3 vs L4-L5	MLkNN	80%
	Abuse (Hate Speech + Offensive Language + Profanity)	L1-L3 vs L4-L5	MLkNN	82%
	Politeness (Apology + Gratitude + Reasoning)	L1-L3 vs L4-L5	MLkNN	93%
	All (Polarity + Hate Speech + Apology)	L1-L3 vs L4-L5	MLkNN	96%

Table 5.4: Classification task performance using TF-IDF and BOW, politeness, abuse, sentiment to predict disagreement. A classifier that uses top features of abuse, sentiment and politeness is able to accurately predict whether a text contains polite or abusive disagreement.

comment contains abusive (L5-L4) or polite disagreement (L1-L3). The following task was to perform a multi-class classification task applying BOW approach in the random forest classifier to predict the disagreement level (L1-L5), reaching low accuracy rate (53%) and when predicting (disagreeable comment (L1-L3) reached (73%). This is mainly caused due to problem of the class imbalance when tagged by the crowd workers. The random forest classifier outperformed logistic regression and other algorithms with an acceptable rate in binary classification reaching F-score of 84%. This is mainly due to the fact that random forest classifier can handle high cases of noisy data with decision trees of each labelled text [241].

The final task of classification as shown in Figure 5.11 was to build a multi-label classifier to test the three main categories abuse, sentiment and politeness to determine the disagreement on a given comment. In particular, using the adapted algorithm MLKNN. The classifier achieved accuracy (80%) when using sentiment features, (82%) when using abusive features and (93%) when using polite features to predict the level of disagreement. Applying the top textual feature from each category including hate speech, polarity and apology was the best to detect disagreement (96%) accuracy as shown in Table 5.4.

Furthermore, the Bayesian Network graphical model ⁶ was used to understand the conditional probability of such variables than between abusive or polite text features pre and post disagreement as shown in Figure 5.12. To apply all factors to binary (yes/no)

⁶<https://cran.r-project.org/web/packages/bnlearn/index.html>

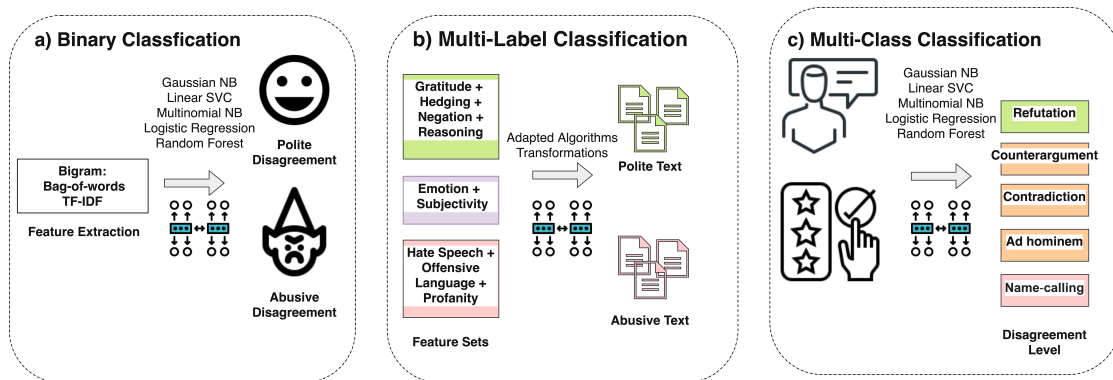


Figure 5.11: Summary of the three experiments, multi-label and multi-class classification tasks.

format for analysis, the levels L1-L3 were converted to polite disagreement (yes) and abusive disagreement L4 and L5 abusive disagreement (no). Also, used average for words count per comment to assign (yes) if it is on or above average and (no) if it is below the average number of words.

The rest of textual factors were already assigned to binary value. Prior disagreement, text that contains negative emotion is most likely to lead to either hedging or negation. Hedging can cause reasoning which lead to reassurance. Negation (N) can case either positive emotion (+) or vote score (S). If positive emotion exists in a text, then gratitude (G) will mostly likely to appear after that.

When disagreement (D) is used, it can impact the length of the comment or profanity. The profanity is escalated by three main factors: hate speech (H_e), offensive language (O) and disagreement. Hate speech can case negative emotion and offensiveness. The number of words used in a comment can affect the sentiment and vote score. The nodes on the right side of Figure 5.12 labelled with red arrows shows the abusive disagreement path and labelled green arrows is polite path . The disagreement node (D) is in principle the outcome that the study was after, which was collected from the crowd workers in Section 5.3.

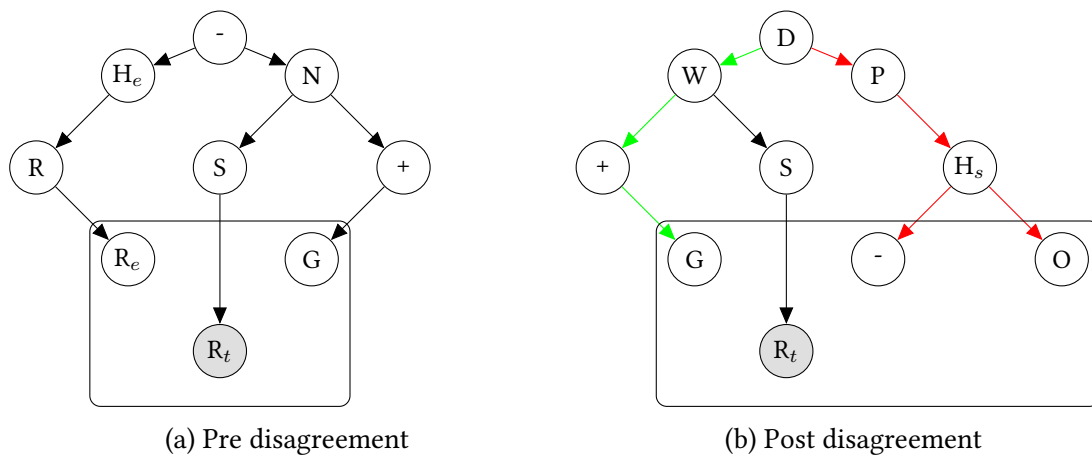


Figure 5.12: Bayesian networks model showing conditional probability between the abusive ↓ and polite ↓ variables before or after disagreement (D). Each node is represented by the initial letter, i.e., node S is score of up-votes - down-votes, D is disagreement and so on. Node H_e is hedges and H_a is hate speech. The '-' node is negative emotion and '+' is positive emotion. In summary, time to respond to a comment R_t is caused by the feedback score obtained from other contributors. The negative emotion can lead to abusive reaction.

5.4.4 Predictive analysis

Interaction effects in regression can show how values of a depended variable that play an effect on independent variables including polite, abusive sentiment textual features. Figure 5.13 shows plots for up votes and response time against selected aggregated features: politeness, abusiveness counts and sentiment score. Politeness counts were calculated as the number of occurrences of hedging, negation and reasoning textual feature. Abusiveness was calculated as the number of occurrences of hate speech and offensiveness. The sentiment score is represented between 1 (positive emotion) and -1 (negative emotion).

The plotted y-axis in Figures 5.13a, 5.13c and 5.13e represents the predicted variable which is the mean up-vote score for the comments of different disagreement levels. The x-axis plots the dependant variable, the mean score of each textural feature set. These suggest that disagreement level on the up-vote score is dependant on the polite, abusive and sentiment textual features. In particular, ad hominem level in all figures for up-vote vs textual features suggests that it is statistically significant. The lines are

not parallel, and thus show that the slope has positive trend based on disagreement level. For example, Figure 5.13a suggests that as the politeness count increases, so does the up-votes score increases. In addition, the up-vote score is greater for comments classified as ad-hominem compared with counterargument and contradiction levels. Abusive comment in Figure 5.13c indicate that the up-vote score increases when a text contains abusive content in counterargument level and significantly more in ad hominem level. There is no effect for abusive comment on up-vote score on contradiction level. Sentiment comment has similar effects of polite disagreement as shown in Figure 5.13e.

We can predict the disagreement on polite comment by measuring time length of reply as shown in the provided Figures 5.13b, 5.13d and 5.13f. Counterargument takes much less time to reply to comment politely, then L3 takes a bit longer but not more than three hours. Yet, L4 takes significantly longer time and reaches the mid point when disagreement is at maximum point in L2 and L3.

We can see that disagreement on abusive features can be predicted from the level based on the the reply duration time in seconds. The results show that ad-hominem level is most likely to commit abusive comments quicker on average time when replying to another comment. Counterargument on the other hand, tend to take longer time to reply or less likely to happen immediately. Contradiction is not affected by time length of reply.

The sentiment comments seem to show positivity quicker when categorised as counterargument comment and similarly in counterargument. It takes longer time to express positive feelings in L4. To test the significance of disagreement against the main categories (abuse and politeness), Johnson-Neyman [242] intervals slope test was used. When Illiteracy is OUTSIDE the interval [1.80, 2.90], the slope of Hedging is $p < .05$. The results revealed that Hedging showed the significance between L3 and L4, and profanity is in L4 significant. The outside range in negation is 1.80,3.20] and $P < 0.5$ in L4.

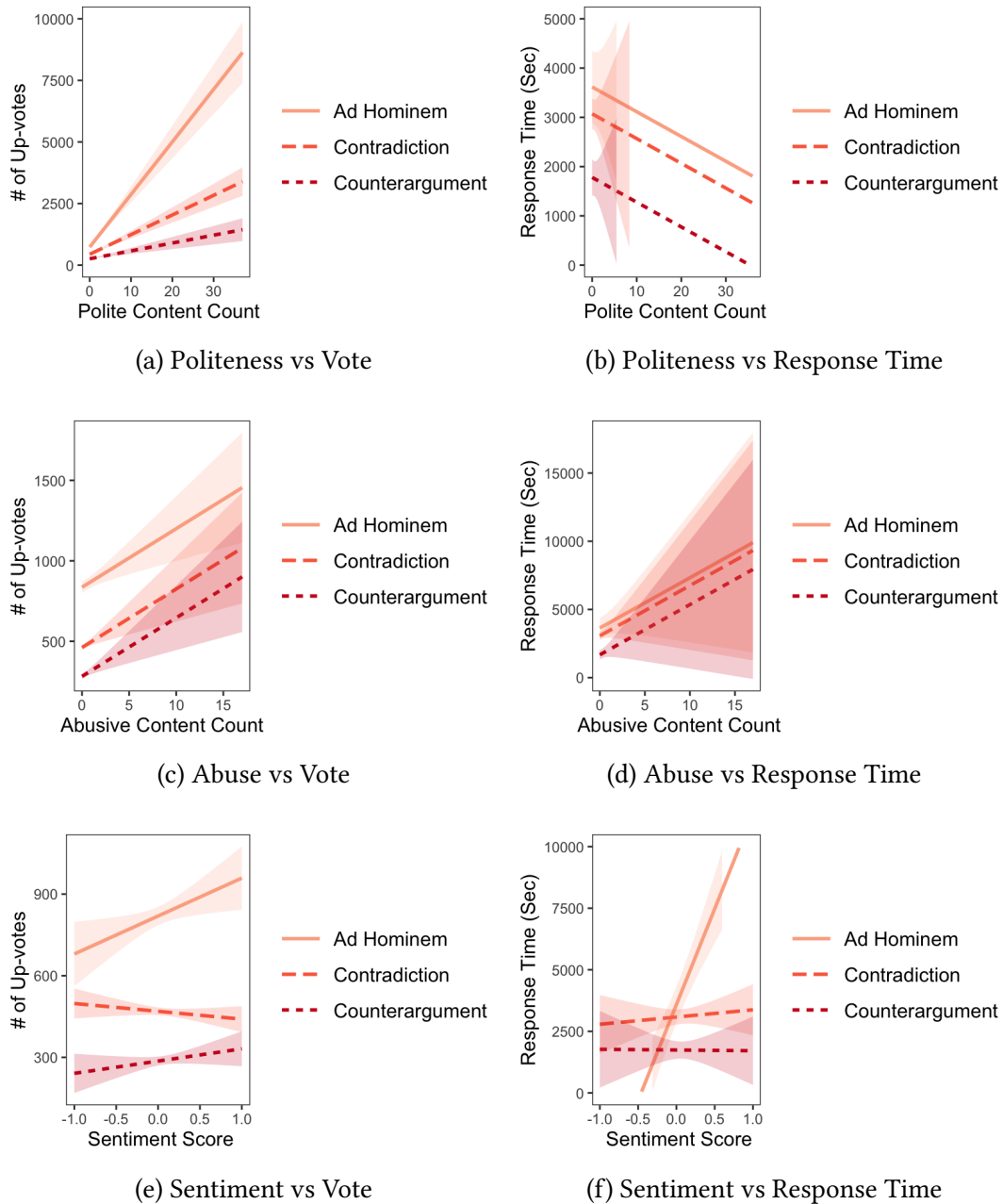


Figure 5.13: The interaction effects using regression analysis on the disagreeable levels ■ Counterargument (L2) , ■ Contradiction (L3) and ■ Ad-hominem (L4) are used to predict score and response time. The figures (a)(c)(e) display the interaction effects between the three categories of disagreement and up-vote score. The figures (b)(d)(f) show the interaction effects against duration time to respond. The relationship between vote score and duration of reply time against textual feature set is affected by disagreement level. Positive slopes in figures (a)(c)(e) and (f) indicate positive interaction relationship between the predictor variables and thus are statistically significant. Negative slopes of all levels in figure (b) and contradiction in figure (d) indicate that they have negative interaction, meaning that when slope value of independent variable (polite, abusive and sentiment disagreement) increases, then response time decreases.

5.5 Discussion

5.5.1 Current design implications

Voting mechanism. Unlike YouTube platform where users gain popularity and get paid by the number of views, redditors engage in an online community while also deciding which posts receive most popularity by up-voting or down-voting accordingly. It can be explained as an online discussion board on which users are given the freedom to pick which content merits to be seen and which doesn't. The number of down-votes minus the number of up-votes determines the post's score, where the posts with the most up-voted scores appear at the top of each page.

Furthermore, Reddit uses a rewarding technique namely Karma points that can be accumulated from the score of votes. [230]. This mechanism, however, can often enable abuse of the system in many ways. To mitigate such problem, an intervention mechanism should allow users to see the scale of abusiveness and receive alert messages to avoid further conflicts between users and moderators.

Gaming the system. To maximise the exposure of a post on Reddit, users' goal should ultimately be to reach the front page [243]. Regardless of whether if it is a bot disguised as a person or an advertising firm disguised as a person, how exactly does a user accomplish that? Is there a particular method to reaching the top? The answer is yes, due to duration of time, submission type, phrasing of the title, buying up-votes, etc., users will all have an effect on the success of a submission. Yet, the average Redditors might argue that the most effective method of accumulating Karma is to copy and re-submit a post to another subreddit.

Gilbert [244] found that within a 17 day period, 52% of the submissions that reached the front page had previously been posted within that same period. This means that 52% of the submissions were re-posts while the other 48% were either original or re-posts that merely fell outside the given time frame. In other words, users more likely to reach the front page by re-posting a pre-existing submission than if they were to post something original.

5.5.2 Theoretical implications

Up-votes and down-votes are essentially connected to how threads and comments are observed. It might be obvious for users to solely down-vote something they disagree with and up-vote elements that promote their opinions. Prior studies have suggested that community feedback can alter or re-shape people's views [72]. This work showed that understanding the factors of abuse and politeness can be captured to better detect disagreement across discussion communities. This can contribute to the moderation intervention system to flag particular users' or groups' comments when they are approaching the borderline of disagreement and abuse.

When a discussion is made and a user decides to comprehend the discussion of a post or comment, it may escalate so quickly due to a subsequent position about the opposite opinion that would lead to shift the argument apart. Yet, it is almost hard to know whether a user is adequate to express a polite or abusive disagreement immediately, which puts the arguer in a fairly defenseless position. Prior work [245] has suggested that factors as interdependence and insecurity play a significant role in terms of trust in online discussion communities.

Most of the comments in the abusive scale in this work expressed forms of anger and insults that attempt to shut the discussion down immediately. On the other hand, the disagreeable scale tends to show more evidence of supporting an argument. This can promote misusing votes based on what should be disagreeable instead.

5.5.3 Limitations

Labelling task. During the final review of the collected data, it was observed that some of the comments were out of boundary scale (not classified as polite nor abusive disagreement). This is mainly due to the fact that there was no choice for non-disagreeable texts during labelling task, i.e, if the comment does not contain a text matches to any of the five levels of disagreement. To mitigate this problem in the analysis, the labelled data was reviewed and tagged as disagreeable comments, then labelled those particular comments, (e.g, "I am making SO MUCH popcorn right now!.") as non-disagreeable of which comprise 379 out of 3690 comments (10.2%). In addition, those comments were

identifiable since they had low rate of agreement score among judders. This problem, however, may also cause class imbalance or/and over-fitting the model if we use all levels rather than the targeted three disagreeable levels.

Sample size. Although the sample size is small and is not intended to be representative of all online discussion platforms, this work does offer a broader perspective of how disagreement and votes in discussions are still a challenge for moderators and moderator systems.

Evaluation. Although a systematic approach was followed to estimate the efforts of the crowd workers, some workers had a hard time to correctly label the comments. This could be owing to the missing details about the entire conversation thread or at least the previous comment to understand the structure of the conversation. However, this is a challenging task since some of the comments were excluded in the samples due to insufficient textual features or nonsense comments. The reward amount or clarity of instructions may also play a significant factor to receive such low level in agreement among raters while labelling the comments.

Another limitation in this chapter is that although consensus amongst crowd worker tasks were measured, no checks were made to ensure that a worker was engaged with the task. This is problematic because workers may reach consensus, even if they are all disengaged. Several authors have noted the problem of high worker disengagement in crowd tasks and proposed countermeasures. To lower the risk of threats to validity of the collected data, several approaches can be taken during the crowdsourcing step. McDonnell et al. [246] proposed an approach comparing rationales amongst annotators in a given task. Annotators provided reasons for selecting one subjective answer, then answers are evaluated by their peers to decide if the justification is valid. The analysis revealed that the rationales can improve transparency and quality of collected annotated data. Also, the analysis suggests that the purposed approach did not impact time of submitting the answer of task, and can reach (96%) of accuracy when combining five workers' responses to discover the correlation from judges' justifications. Rzeszotarski and Kittur [247] used behaviour fingerprinting to measured the behaviour of crowd workers while doing the task to determine the quality of submission. The behavioural actions and timing include number of key-presses or clicks and duration of the task. The results showed that the approach can help identify low-quality or unacceptable

submission.

For the current work, these approaches were not feasible due to resource constraints. The consensus measure across all tasks showed that the α reached to an acceptable reliability. This indicates that although there is a risk of disengaged tasks affecting results, the risk is acceptably low.

5.6 Summary

Having analysed the selected sample of comments, the analysis revealed that negation and reasoning textual features can help classify the level of disagreement, in this case between contradiction and counterexample. Hate speech and offensive language were mostly used in the contradiction followed by counterargument levels. In all polite features, counterargument (L2) is mostly used in conversations. Hedges decrease mostly in L3 when L4 disagreement is used. Reasoning is almost linear in L3 and L4 and low. It is typical to see negation is significantly higher in L2. Users tend to overall use less gratitude words in conversations. It is so evident that abusive words in comments were mostly used in L4, e.g. profanity and offensive languages and hate speech. In most cases, L3 and L2 intersect during 10 am and when the number of abusive textual features in L2 begin to grow, the number of abusive behaviour tend to decline in L3. Negative emotions are mostly shown in L4, while positive emotion is mostly spreading in L2. These findings confirm that the disagreement tends to reach the borderline of abusive level in the disagreement scale. When an argument is in the refutation level, it tends to show strong positive and negative feelings and avoid profanity. Also, the text is more readable and subjective. This could be due to the fact that these comments at this level contain a larger number of words. The analysis shows that vote abuse can often be caused by disagreement, particularly when it affects reputation of the user by losing or earning points. As shown in Table 5.3, refutation-level comment is scored -2 and 15 on an ad hominem level example. This could be due to a user lack of understanding of the guidelines, or antisocial or mental problems, situational or competitiveness factors [108].

The disagreement arises from both abuse and politeness factors—whereas previous work has focused on a physiological perspective, the presented approach by contrast,

has used textual analysis to show that both politeness and abuse context affect disagreement behaviour. This suggests the significance of different design affordances to manage either polite or abusive disagreement. Rather than removing or deleting all comments which are abusive and break community guidelines, considering measures that mitigate the disagreement factors that lead to abuse may adequately speculate the tone of conversation.

RQ5. What kind of context enables and promotes polite or abusive disagreement on an online discussion? Do particular kinds of disagreement trigger down voting?

Overall, the RQ5 was employed to find how disagreement is predicable based on the purposed five measures of disagreement. The analysis revealed that disagreement detection can achieve a high accuracy rate when considering abuse and politeness features in a given text. Understanding the disagreement levels in an argument can help to see the evolution of a conversation or discussion which may lead to reflexively attacking the character of the user who is making an argument rather than the body of the argument itself. Also, provides an understanding of how disagreement can impact vote.

Chapter 6

The Interplay between Disagreement, Abuse and Moderation in Online Discussions

6.1 Introduction

Abusive content in online social platforms discourages contributors to stay in the community and may create social conflicts between contributors and moderators. Existing approaches to moderation focus either on manual detection, or on the detection of abusive behaviour in isolation from context, leading to overburdening of moderators and reduced quality of discussion. Thus, the design of social platforms require developers to rethink intervention strategies and improve moderation systems. In particular, we presented analyses about the interplay of disagreement and abuse in online discussions that impact moderation decision making. These contributions imply that moderators have a tendency to react strictly against contributors to protect their communities, as intervention occurs late in the discussion. For example, they may ban a user who violates community norms and guidelines without understanding the factors of abuse. This may not be the best approach to tackle such a problem. Moderators rather need to understand the causes of abusiveness that originate from disagreement and further how the online setting can accelerate the development of abusive behaviours. If mod-

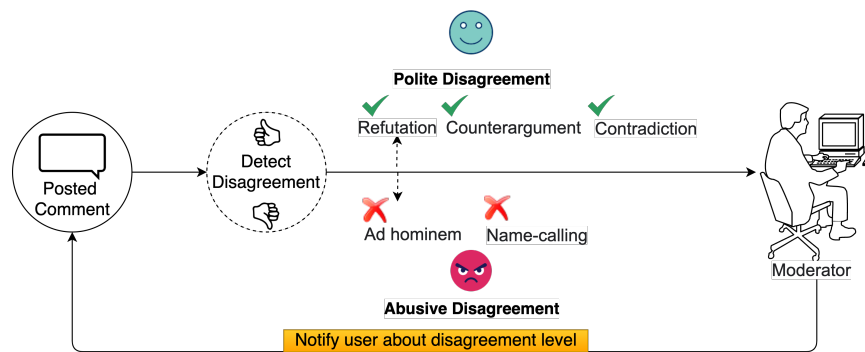


Figure 6.1: Summary of contributions that suggest an anticipatory moderation system design to combat abusive content and disagreement. The comment is posted by a contributor to identify features that capture the disagreement level to classify comment based on the category level of polite and abusive disagreement. The moderator then reviews issue about disagreement and notify contributor that discussion may lead to abusiveness before intervention.

erators can distinguish between polite and abusive disagreement, then they can inform or alert contributors before contributions become abusive. In this chapter we propose an anticipatory approach to moderation based on the contributions in earlier chapters that encourages users to interact less aggressively.

Figure 6.1 represents a summary of the anticipatory interactive moderation system design that has been studied in this thesis. The suggested approach in Figure 6.1 shows how moderators can interact with contributors by sending notification message about the fact that the discussion may lead to abusive content due to disagreement. The figure illustrates how, before moderators are reviewing the comments, the classifier identifies the level of disagreement to assist human-decision in intervention strategy. The analyses in Chapter 4 showed that online and in-person comments differences can help capture abusive behaviour. In particular, online comments showed strong negative behaviour while in-person comments contain greater polite terms. This is essential to train a classifier to predict comments that may need to be removed using particular textual features including reasoning, reassurance, gratitude, apology, swearing and offensive language as shown in Figure 4.10. Therefore based on this analysis, the system presents contributors with the opportunity to intervene before the type of disagreement within a conversation declines as a whole.

Each following section reviews how each individual contribution fits into the overall system. In each case, further relevant literature is also discussed.

6.2 Moderation anticipatory system

To build an interactive anticipatory system for content moderation, we need to understand the interplay factors between moderators and contributors when reacting to a particular event. This section discusses how moderators react as facilitators of good disagreement, rather than enforcers or punishers of abusive disagreement, and how the results of Chapter 3 support the case for anticipatory moderation system which fits to the big picture of this research.

The first explored factor is to measure the activity of moderators and contributors. Particularly, Figure 3.4 showed in Figure (a) the total number of hourly submitted comments was mainly during working hours in the morning, while in Figure (b) we showed that the number of intervention declines during the week days in the morning time. In Figure (c), we showed that controversial comments tend to appear mostly during morning hours and less likely in late nights. These findings suggest that controversial comments are more likely to appear when contributors are more active to post comments and moderators are less active to intervene in most cases which indicates that there is a lack of communication between contributors and moderators. So, moderators need to learn how to approach contributors that are about to commit abusive content to reduce the possibility of enforcement banning a user. Next, we discuss some evidences from the relevant work that are connected to the observations found in this work to address the need for anticipatory system for content moderation.

Moderators have multiple intervention strategies against abusive behaviour to take a particular action anything from removing a comment to banning a user account. In most cases, moderators may have to remove a posted comment without a review due to the massive received number of submitted comments and less number of available moderators. In Chapter 3, we showed that the number of intervention activities can impact the number of reviewed posts as seen in Figure 3.3b. In particular, the number of intervened posts declines as the number of reviewed posts increases. This suggests moderators need to use tools that assist them during the review process. The

anticipatory moderation system allows moderators to review in less time because the moderators will be able to identify the level disagreement and ask contributors to be cautious about the language before intervention.

Another issue of content moderation system relates to the background perspective. We showed in Table 3.2 that most male and younger moderators (18-29) take removal action followed by ban then suspend actions. Female moderators are significantly lower than male moderators when it comes to escalating the problem to higher authority and banning a user account, and more are less likely to suspend accounts. Cheng et al. [248] surveyed 20K Facebook users to investigate the core factors of background perceptions related to the social media abuse. The authors claimed that younger male users are most likely to experience online abuse and lack of control of their behaviour due to excessive use of the social media. The authors suggested a development of on-line interactive tool that can help users to control their behaviour. These observations suggest that age profile, gender differences and nationality background play a significant factor to impact human interactions between moderators and contributors. For example, one moderators may justify one reason for post removal is that it lacks a clear argument or leading to an assault which is breaking the communality rules. On the hand, a contributor may respond that the post contains a typical argument and can not understand why the moderator decided to remove the post. So, implementing an interactive anticipatory moderation approach may promote positive interactions and facilitate in eliminating the conflicts between moderators and contributors that are originated from perspective differences.

Prior work has investigated the factors of spreading online harassment and hate speech and how moderators respond to them. Chandrasekharan et al. [69] examined two communities in Reddit namely r/CoonTown and r/fatpeoplehate and showed that contributors that use hate speech can inherent factors of abusive behaviour to spread their behaviour to other similar communities. This behaviour can lead to social movement or extremism which requires extensive cases of interventions. In Chapter 3, one moderator in section 3.5.5 addressed that some contributors post pictures or expressions that can lead to abusive behaviour followed by spreading social movement against particular group. We also have seen that rule breaking and irrelevant content are the primary reasons for intervention as reported in Figure 3.6. So, reasons for interven-

tion and activity are significant factors that can help to identify abusive behaviour in online social platforms. In particular, users tend to influence one another over time across multiple online communities which means if a contributor breaks the community guidelines or norms, s/he is mostly likely to influence their behaviour to another fellow contributor. However, the anticipatory moderation approach is encouraging moderators to communicate with contributors to build a reliable relationship which influence users to adjust their behaviour.

Several authors have investigated most common reasons for moderators to intervene. For example, Jhaver et al. conducted a qualitative and quantitative study of more 900 users on Reddit to investigate the reasons form post removal from users' perspective [249]. The analysis revealed that users who are aware of the guidelines and understood the justifications of post removal from moderators are more likely to stay in the community and post again. In Chapter 3, we found that the most common reason for intervention reported by moderators is violating the community rules as shown in Figure 3.6b. This confirms that community guidelines play a significant role in the removal process in content moderation. Nerveless, contributors may feel that they were dismissed or banned from the community without a valid reason for their behaviour. So, an anticipatory approach allows moderators to be more explicit about the removal reasons and avoid the spread of abusive behaviour because contributors may leave the community but contribute abusive content to another communities since they did not have the opportunity to engage in constructive understanding of their behaviour

Several authors attempted to present assistive tools for content moderation. Chandrasekharan et al. [30] developed a moderation assistant system to help moderators review comments in order to capture abusive behaviour. The approach relied on cross-moderator decision to take particular action. For example, if moderators reach 90% of agreement, the comment is most likely to be appropriate for removal. However, this this approach can be difficult to apply learning when different communities adapt multiple rules that is ordinary in one community and yet uncommon to another community. The approach also resulted in lower accuracy rate (86%). Reviewing less number of comments resulted in more number of interventions as reported by moderators in Chapter 3. This indicates that moderation can be a time consuming task. Anticipatory content moderation may help mitigate this by considering the overall content of

a conversation. The proposed approach gives moderators less time to review posted comments and less number of cases for banning accounts.

Lastly, Chandrasekharan et al. [36] have shown that abusive behaviour on Reddit most likely leads to racism or other forms of discrimination against religion, ethnicity, gender and political views when discussing conversational topics. As seen in Chapter 3, we showed that the most intervened topics were in politics followed by gender and religion, as can be seen from Figure 3.7. In addition, one active moderator reported in section 3.5.6 that topics related to Donald Trump or Islam religion it leads to personal attacks or direct insults. This suggests that the approach of anticipatory moderation requires moderators to spend more time to make interactive actions in debatable topics. In particular, most conflicts about intervention were related to disagreement about particular point of view, which can escalate from legitimate disagreement to abusive content.

6.3 Online and in-person differences and similarities

This section discusses the contribution of interplay factors of online and in-person differences in conversation reported in Chapter 4 that is related to the design for moderation anticipatory system. To build a better interactive moderation system, we must understand how contributors behave in different settings of online discussions. In particular, the differences between both settings discovered in Chapter 4 shed the light on selected textual features to capture online abusive behaviour in discussions. The results also showed how consensus building factors between the two settings are different. For example, online conversations tend to show less use of hedging language and being more subjective. The analysis contributes to online abuse detection which is a critical element for the design of anticipatory moderation system.

In Chapter 4, we showed that the textual detection approaches targeted multiple polite, abusive and sentiment features to predict, setting of conversation, removed comments and disagreement level. In addition, we have seen that online and in-person conversations detector can perform better when considering polite and abusive content as follows: reasoning, reassurance, gratitude, apology, offensiveness and profanity textual features. In particular, combining highest scores of predicting if the comment is most

likely to be abusive or polite between both in-person and online settings. Furthermore, we showed that the occurrence of profanity with reassurance, gratitude and apology textual features have improved the performance for predicting abusive content. In the literature, Sood et al. [137] trained labelled data by MTurk workers and claimed that the existing list-based approach is less effective to capture profane terms in the context. The authors suggested that combining Levenshtein Edit Distance and SVM with list-based approach can improve the performance for predicting profanity. Ultimately, we want to be able to build a detection tool that can allow moderators review submissions to decide whether a submitted comment embodies abusive or non-abusive content in terms of politeness features measured by both settings.

Effective interactions with positive feedback amongst contributors can promote contributors to behave politely and encourage moderators to become more tolerant. Cheng et al. [27] claimed that contributors who submit fewer posts in online discussion communities are those whose submitted their fellow contributors' posts and/or off-topic content. Also, contributors whose were banned from the community received harsh feedback and submitted low-quality content over time. Another work [31] found that negative feedback can promote negative behaviour. Our findings are in-line with the findings of [27] and [31]. For example, online discussions were mostly off-topic in comparison with in-person discussions. Also, online discussion showed more subjective terms in the context than in-person, but subjectivity score declines over time. This may imply that people are more confident to express ideas. Therefore, they proposed system of anticipatory moderation can prompt contributors and moderators to interact effectively, rather than completely banning the users for particular behaviour.

Online or in-person communication can influence the outcome of conversation. As examined in 4 the two settings of conversation in-person and online (asynchronous) were different in Figure 4.7. In particular, online communication showed extreme sentiment text. In addition, online communication tend to use more subjective terms and easier to read. These findings can suggest that in-person communication is less likely to include abusive content or stay off-topic in the conversation. Thus, it suggests that interactive approach would be effective in keeping the discussion safe amongst contributors without having extensive number of interventions by moderators.

The analysis performed in Chapter 4 is crucial to distinguish between behaviour in

online and in-person in order to unfold the trajectory of abusiveness in discussions. As reported in Table 4.3, all proposed textual features except negation have rejected the null hypotheses which implied that they are valid features to apply in disagreement detection for implementing the anticipatory moderation system.

6.4 Detecting abuse and disagreement

This section discusses the reported findings in Chapter 5 about the interplay factors of disagreement and abuse detection mechanism to facilitate the anticipatory moderation system and compare it the literature. As seen in Chapter 3 and Chapter 4, users behave differently in online conversations based on variety of reasons. There are two distinct detecting features, either by activity of users or textual features. In this thesis, we only explored activity features and dedicated our analysis on textual features to reveal the differences between disagreement and abuse in a comment. Also, to detect the type of disagreement that contains polite or abusive content. These analyses are essential to improve the engagement between moderators and contributors in the proposed moderation anticipatory system for online communities.

Most conversational methods for detecting abusive content rely on conventional resources of Natural Language Processing (NLP). For example, Nobata et al. [135] developed an abusive content detector using linguistic, syntactic, N-gram, semantics features. Their detector model that combines all proposed features reached (78.3%) of F-score. In Chapter 5, we presented five levels of disagreement to differentiate between legitimate and abusive disagreement. The scale contains five levels: refutation (L1), counterargument (L2), contradiction (L3), ad hominem (L4), name-calling (L5). The first three levels are categorised as polite disagreement and the last two levels are categorised as abusive disagreement. To better identify the level of disagreement, we classified each level by extracting three main textual features: hate speech, polarity and apology. Our disagreement metric that uses three characteristics of textual features have reached 84% in F-score and 96% of accuracy rate in multi-label classification tasks as summarised in Table 5.4. This shows that disagreement measure is an essential element to be able to find the differences between polite and abusive disagreement so that moderators can inform the contributors if they reach to high disagreeable level of

(L3 or L4).

Identifying online abusive behaviour can become a challenging task due to lack of resources of textual features to build a better classification model to capture abuse. Davidson et al. [25] have proposed an abusive content classifier to capture hate speech and offensive language using crowdsourcing to label assigned tweets. However, the classifier did not perform well in detecting hate speech reaching 61% accuracy rate. The authors indicated that one possible problem is that there are some instances where hate speech was misclassified due to the lack of use of profanity or possibly crowd workers misunderstood the difference between hate speech and offensive language. Our analysis in Chapter 5 showed that disagreement can lead to profanity in Figure 5.12 which is caused by hate speech and offensive language. Hate speech is caused by negative emotion. This implies that negative emotion is a key factor to lead to negative behaviour and disagreement. So, it is important to detect disagreement at a positive level before it escalates from polite to abusive content in order to keep the conversation constructive. Moderators can benefit from this anticipatory approach to decide whether the submitted comment falls into disagreeable or abusive level.

Voting system can be complex and misused or misinterpreted by contributors. Most discussion platforms design buttons to allow users to interact to particular content, i.e., dis/like or up/down-vote. For example, Reddit platform defines the up-vote button to indicate whether the comment is relevant to the subreddit or not. Such facilities can involve contributors as collective moderators, if these ratings are used to inform or direct moderation. However, this may not be appropriate, since a contributor's reason for rating a comment may concern the extent to which they agree or disagree with it, rather than primarily judging its relevance to the debate. In particular, contributors may misuse the votes by down-voting comments about an argument from point of view because they disagree with the claimed statement. This can cause comments to become more controversial due to votes conflict and become hard to distinguish between dislike of and disagreement with a comment. Jhaver et al. [249] stated that users who are spending time to improve the quality of their post are less likely to post again because they feel that the reason of removal is not fair. As we have shown in Figure 5.13a that as the number of polite words increases in a single comment, the response time decreases based on the disagreement level. Figure 5.13d showed that

response time increases when the number of abusive words is high in disagreement. Also, Figure 5.13f showed that it takes more time to reply for L4 comment that contains negative emotion. Having a precise disagreement detection tool can help understand the key issues related to vote abuse and disagreement as suggested in the anticipatory moderation system.

Similarly, the the opinions in the sub comments listed in each conversation thread in the same post will also be down-voted due to disagreement. In particular, we have shown in Figure 5.9a that L4 disagreement tends to contain higher score of subjective score over time and we know that the number of up-votes decreases if the comment includes abusive textual features in disagreement L4 as shown in Figure 5.13c. Additionally, most social platforms sort up-voted posts or comments at the top of the page to be seen first as a default set up. As seen in our vote and disagreement findings listed in Table 5.3 and Figure 5.13, votes were mostly affected by disagreement level. In particular, the example table showed that a comment that was in the refutation level received negative votes. Also, Figure 5.13a, 5.13c and 5.13e showed how disagreement level can impact the up-votes based on comments that contain polite, abusive and sentiment features. For example, Figure 5.13e showed that negative comments lead to low number of up-votes in L4 disagreement and positive comments lead to high number of up-votes in L3 and L2. Therefore, It is much simpler if moderators were able to identify this kind of issues related to vote abuse earlier to avoid further implications. As shown in the proposed anticipatory moderation system in Figure 4.10, detecting disagreement can help moderators identify the level of disagreement which is often misguided with votes.

Finally, automated approaches that are based on votes may discard many posts which discourage new contributors to post frequently. In other words, contributors may feel less interested in posting and sharing their ideas if they observe many down-votes on the subreddit that they like that can harm their reputation if they decide to post any content that resulted in many down-votes. So if the automated system is trained to detect abusive comments based on community vote, it may remove comments that are in the polite disagreement category. Thus, we must consider implementing anticipatory moderation system to eliminate vote conflicts between moderators and contributors.

6.5 Summary

In this thesis, we have presented novel work concerning the characterisation of abusive behaviour in online social media that contributed to both computational social science and social computing research. In particular, the work investigates the key reasons for spreading abusive comments and how both moderators and contributors react to particular actions or interventions. In addition, this thesis examines several textual features to detect abusive content. The following are the key findings of this work:

1. Most moderators reported that their role is a member who is trying to preserve community's values, and fewer moderators reported that their role is an employee to take actions of moderation as shown in Figure 3.2. As a result, Figure 3.4 showed that contributors tend to become more active during working hours while moderators are less active.
2. Whenever moderators take time to review posted comments, then it requires a less number of interventions as shown in Figure 3.3b.
3. The differences between online and in-person conversation in terms of hedging, hate speech, offensiveness, cosine similarity, readability, sentiment and subjectivity textual features are statistically significant as shown in Table 4.7.
4. As shown in Table 4.7 combining profanity, reassurance, gratitude and apology textual features showed high performance of accuracy to predict comments that will be removed by moderators on the Reddit platform.
5. The qualitative analysis in Figure 4.11 showed most frequent factors of consensus building in online and in-person comments. For example, in-person comments showed a degree of consensus building in terms of agreement or disagreement.
6. The disagreement levels model can be used to distinguish between polite and abusive disagreement. Table 5.4 showed that having polarity, hate speech and apology textual features is able to identify the disagreement level.
7. Disagreement plays a significant role to impact up-votes and response time. For example, Figure 5.13a showed that a comment that contains less polite features in

the ad-hominem level (L4) declines the number of up-votes. Also, Figure 5.13f showed that it takes more time to reply in L4 when the comment contains negative emotion.

These findings are important because they represent suggestions to combat online abusive behaviour by proposing an interactive and anticipatory moderation system in social platforms to make the moderation process much simpler and to reduce the risk of discussions becoming abusive.

User-generated content (UGC) is a key aspect in the research area social computing. Social computing permits people to communicate much simply with one another. Additionally it can promote in-person interactions between many organisations though social platforms. In Chapter 5, we showed that detecting disagreement, investigating the differences between conversational settings in Chapter 4 and understanding the perception of content moderation in Chapter 3 are more effective methods for identifying abusive behaviour than previous approaches from the literature, e.g., [27], [25] and [36]. In particular, as shown in Table 5.4, our study showed that combining sentiment, hate speech and apology textual features can effectively identify polite and abusive disagreement. These approaches which contribute to the field of computational social science and social computing suggest re-designing intervention strategies to early capture polite and abusive disagreement, and to ease any conflicts between contributors and moderators.

In this work, it was evident that disagreement is one key factor in conversation that can lead to abusive content. So, implementing features that can distinguish between polite and abusive disagreement as proposed in this thesis can help moderators to take intervention actions simpler, which also can be misled with up/down-vote.

Exploring characterisations of abusive behaviour in online settings has shown how moderators and contributors behave differently online. In particular, the analysis including the selected textual features to capture disagreement and abusive content showed improvement in detecting conversations containing abusive contributions. All these tools and approaches are helpful to social computing research area that is looking to reconstruct social conventions across multiple social platforms. The work creates the potential to assist moderators in all online discussion communities. This can allow users to engage in safer and more productive environment.

The thesis also contributes to the areas of computational social science (CSS) and soft security (SS). The proposed methods of extracting automated information and social media analysis are useful to the CSS research community. In addition, the contribution of the thesis applies approaches that identify abusive behaviour related to the SS domain in terms of studying the social norms to build trusted moderation system in peer-communication platforms.

Chapter 7

Conclusion

Abusive behaviour is a key issue on any social media that can prompt significant impacts and affect the reliability and quality of posted content on an online community. The form of abuse includes hate speech, profanity and offensive language. The setting also of online conversations through social network platforms offers an environment in which abuse can be problematic. Moderators of online communities additionally react differently in explicit strategies to protect the values of such community. The vast majority of these platforms rely on traditional methods to detect abusive behaviour. Not always online comments are removed due to abusiveness, disagreement can cause people to react differently and post abusive content. Thus, there is a need to investigate the causes of abusive behaviour in social media settings, in particular, the disagreement in terms of politeness and abusiveness. Also, understand the moderator' behaviour and perspective.

In the previous chapters, research which shaped the thesis aims was outlined and the research questions. To begin the investigation, a survey of moderators along with empirical analysis of data in chapter 3 examined association procedures which permit users to address intervention approaches and perceptions. Chapters 4 and 5 considered textual features for disagreement and disinhibition investigations, individually, showing their possible adequacy and contributing a superior comprehension of how these textual features interplay the degree of online discussion. They additionally explained the later use regarding these modalities for helping users address factors of online abuse. This work in turn found that users or social platform designers could

learn about all together disagreement, disinhibition and perceptions of moderation.

Now, this chapter returns to review the thesis statement that raised research questions and key aims, and mention the potential for future work arising from the findings.

7.1 Review Thesis Statement and Research Questions

This thesis statement was stated in Section 1.3:

This thesis asserts that:

Contributions to online discussions can be detected by classifying contributions in terms of the form of disagreement that they embody.

Abusive behaviour is contextual and may be conflated by community participants, with disagreeable or controversial contributions (e.g. through down-votes), exacerbating the workload of community moderators. Further, online behaviour can be shown to be quantitatively more prone to disagreement and abusive behaviour than in-person, due to the lack of wider social cues and ‘guard rails’. Finally, these insights allow us to classify behaviour in terms of the form of disagreement, distinguishing between polite and abusive disagreement.

Consequently, this thesis explored the occurrence of disagreement context online and how it factors people to behave differently in different settings of conversation. These issues are now explored in more detail by reviewing research questions RQ1-RQ5.

7.1.1 RQ1 and RQ2

RQ1. What role do moderators perceive for themselves on Reddit discussions?

RQ2. When, how and why do moderators intervene on discussions on Reddit?

The relationship between the social understandings of norms, the development of social network site approaches and their community guidelines on moderators is under-researched. Key to this relationship is the role of the moderator and contributor that

shape the online behaviour. To start to investigate this point of view on chapter 3, A study of moderators perceptions on community-based platform namely Reddit was embarked. The findings report that moderators on Reddit see themselves as individuals that belong to their community values and norms. Also, moderators are making a solid effort to provide protection against abusers and encourage contributors to provide higher quality of content. Yet, contributors tend to post most frequently during day nights. In particular, contributors become more active at various occasions when moderators are less active. In addition, moderators tend to intervene or take an action less likely when the number of reviewed posts increases. These discoveries contribute to social computing research community to consider the adjustment for the structure of balance framework that develops moderation systems.

7.1.2 RQ3 and RQ4

RQ3. Is there a statistically significant difference between online and in-person discussions in terms of polite or abusive language used? Can conversation settings be detected?

RQ4. To what extent can stimulated behaviour shape the understanding and perceptions of peer-group evaluation and consensus in discussions?

In chapter 4, it noted that conversations that took a place in-person differ from online conversation in multiple ways. Firstly, the results suggest that variables of disinhibition can help distinguish unwanted behaviour in peer-group learning environment. Specifically, online conversation setting cultivates user practices (less supporting, more outrageous slant, more prominent readiness to communicate sincere belief and wandering from subject) that are known to improve the extreme of harsh behaviour. The classification model can precisely detect online and in-person discussions on the basis of linguistics factors compared to moderated removed comments. All null hypotheses were rejected about the differences between in-person and online conversations except negation language. In particular, politeness, abuse, sentiment and post activities are not the same. The qualitative analysis revealed that users are more likely to comprehend (dis)agreement among their peers regarding misuse. Furthermore, consensus

building variables were shown how they can impact conversations in various settings. Thus, the chapter concluded that the settings of online discussion platforms are more likely to witness abusive behaviour.

7.1.3 RQ5

RQ5. What kind of context enables and promotes polite or abusive disagreement on an online discussion? Do particular kinds of disagreement trigger down voting?

Disagreement often includes various phases of argument. Controversial topics in an online discussion platform can prompt frail reactions or absence of safe and healthy communications. Specifically, confronting abusive behaviour in online conversations. Also, vote misuse can create conflicts between contributors and moderators which can bring about affecting the quality of content and trustworthiness of such an online community. In chapter 5, the analysis looked at 5k comments from top five subreddits on Reddit platform. A disagreement measure was proposed and evaluated that can show the differences between the abusive and polite disagreement into five different levels. Disagreement was shown how it is captured and it can influence votes and discussion. The findings suggested that disagreement can be captured better when including abuse and politeness textual features.

All the research questions that were addressed in this thesis shaped the analyses and findings of such phenomena about abuse and disagreement in particular. Each chapter examined the research questions in different ways. For example, the first experiment in chapter 3 revealed the perceptions of activities of contributors and reactions of moderators. Then, followed by chapter 4 that looked at conversations between peers in group project online and in-person. Finally, in chapter 5 asked whether we can detect polite/abusive disagreement in a context and learn about the causes consequences that led people to behave differently.

7.2 Future Directions

7.2.1 Possible effects of abusive behaviour in discussions

Schadenfreude. Refers to expressing happy feelings towards particular person or action as result of humiliation or failure. Prior work has observed that there are reasons behind schadenfreude which includes aggression, competition, and regulation [250]. These findings confirm the assumption and analysis that hate speech, offensive and negative emotion are valid measures of abusive behaviour. Feather and Sherman [251] reported that displeasure can be identified before accomplishment more than sympathy and envy factors. Additionally, people who are assessed to be undeserving of their victory result in more schadenfreude, when disappointment occurs to them than people who are deserving – despite the fact that they created their own disappointment or not.

Gender conflicts. In the experiment in chapter 4, the majority of participants were males, females significantly were far active in online discussions (77.3%). Females tend to show less positive feelings and subjectivity in both settings. Hate speech and profanity were frequently used by female groups. In terms of politeness including hedging and negation, this was more prevalent amongst males in this sample. This preliminary analysis reveals interesting problems to investigate e.g., gender gap in both settings.

Anonymity. One of the assumed benefits of the Internet as a mechanism for communication is that people are not forced to disclose features of their in-person identity except if they willingly do so. It has been proposed that the anonymity of the Internet can provide possibilities for open-speech due to the fact that people can say what they believe without having any concern that other people will behave or reply negatively clearly due to the cultural belief and background or gender identification [252]. The study did not consider including this feature in the online discussion platform since the team members were allocated based on the group allocation and each know who are the team members. However, this would be an important aspect to investigate in future work.

Group homogeneity. The diversity of groupings is perfect for serving striving team members to learn from a stronger member. In the sampled groups, students tend to

defer to the actual leaders and energetic students to guide the group and perhaps do most of the work. Uniform groupings, on the other hand, can encourage students who did not have the opportunity to participate or contribute enough, which would result in an effective online and in-person discussion. An effect that confers sharing values and goals in each team is an essential key aspect to build a stronger team. This may relate to the understanding of members of an out group as existing homogeneous.

Public vs private. Bridging the gap between private and public conversations is a key factor in understanding how a conversation is held in all settings. In particular, there are elements of conversation and in particular disclosure of personal information that individuals would be happy to share online but not face-to-face, but also conversely, there are elements of conversation that individuals would only share in face-to-face settings. To build a more effective intervention system, online discussion platforms that offer instant messaging in a private setting and public conversation can observe both settings of groups by identifying the measures of abuse and politeness and alert users or groups beforehand if reaching a borderline of abusive behaviour.

Further investigation could consider whether different apps hold similar behaviours between each mode using text analysis. Given that moderators were recruited from one platform, it would be interesting to compare with moderators from other platforms that support different moderation approaches. Furthermore, this thesis highlighted only view textual features based on the initial investigation, yet several text engineering features including intention (e.g. if the text contains complaint, appreciation or suggestion), sarcasm and feelings would potentially reveal more about factors of abusive behaviour in conversational mode. Also, it would be interesting to see some audio or emoji analysis and measure the depression stages in both settings to see how this can affect discussion behaviour. Finally, the proposed labelling approach for measuring disagreement levels in conversations can be used with large-scale data to serve various research disciplines in computational social science and psychology.

Future work may also consider other textual features such as feelings or intention (e.g. sarcasm), and use large-scale data. Community guidelines are also a key factor to correlate each and see how it evolves and how users behave across sub-communities. Some discussion and social platforms including Reddit use a built-in feature namely controversially of a post/comment, which may help to understand issues related to

vote scores and compare it with the result of the text analysis.

Overall, further research can be suggested to continue exploring the perspectives of moderators. In particular, this could include undertaking focus groups with contributors, moderators and mixtures of both groups to further illuminate some of the issues identified in the survey. This should also allow extending the empirical study of moderator's behaviour to understand how and when intervention strategies are applied in different discussion communities.

7.3 Concluding Remark

The primary objective of this thesis is to manifest that abusive behaviour in online platforms is not a reflection of social behaviour in the person (i.e. with fewer constraints online than in person). The medium of online discussions via social media platforms creates a setting that is more fertile and promotes the development of abusive behaviour that also escalates rapidly. Moderators of online discussions also respond differently, engaging in specific tactics to maintain debate quality that are specific to online settings.

From a broader perspective, social platforms designers may need to rethink about principles of online abusive behaviour. This is crucial to combat such behaviour by developing better guidelines, interventions and moderation systems.

Bibliography

- [1] Bailey Poland. *Haters: Harassment, abuse, and violence online*. U of Nebraska Press, 2016.
- [2] Reddit. Reddit by the numbers, 2019. URL <https://www.redditinc.com>. Accessed: 31 August 2020.
- [3] Nancy Willard. Cyberbullying and cyberthreats. effectively managing internet use risks in schools. *Center for Safe and Responsible Use of the Internet. Recuperado el*, 20, 2006.
- [4] Emily Christofides, Amy Muise, and Serge Desmarais. Risky disclosures on facebook: The effect of having a bad experience on online behavior. *Journal of adolescent research*, 27(6):714–731, 2012.
- [5] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.
- [6] Sarah T Roberts. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press, 2019.
- [7] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. Searching for safety online: Managing "trolling" in a feminist forum. *The information society*, 18(5):371–384, 2002.
- [8] Bryn Alexander Coles and Melanie West. Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, 60:233–244, 2016.

- [9] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *CSCW*, pages 1231–1245. ACM, 2017.
- [10] Kathleen Searles, Sophie Spencer, and Adaobi Duru. Don't read the comments: the effects of abusive comments on perceptions of women authors' credibility. *Information, Communication & Society*, 23(7):947–962, 2020.
- [11] Noam Lapidot-Lefler and Azy Barak. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2):434–443, 2012.
- [12] Tim Squirrell. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media Soc.*, 21(9), 2019.
- [13] Christian Fuchs. Hacktivism and contemporary politics. *Social media, politics and the state: Protests, revolutions, riots, crime and policing in the age of Facebook, Twitter and YouTube*, pages 88–106, 2014.
- [14] Michael S. Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G. Vargas. 4chan and /b/: An analysis of anonymity and ephemerality in a large online community. In *ICWSM*. The AAAI Press, 2011.
- [15] Tom Sorell. Human rights and hacktivism: the cases of wikileaks and anonymous. *Journal of Human Rights Practice*, 7(3):391–410, 2015.
- [16] John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.
- [17] James Q Wilson and George L Kelling. Broken windows. *Atlantic monthly*, 249(3):29–38, 1982.
- [18] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102, 2014.
- [19] Pnina Shachaf and Noriko Hara. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370, 2010.

- [20] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*, 2017.
- [21] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*, 2018.
- [22] Ziqi Zhang, David Robinson, and Jonathan Tepper. Hate speech detection using a convolution-lstm based deep neural network. *ESWC 2018: The semantic web*, 2018.
- [23] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. Textual aggression detection through deep learning. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 177–187, 2018.
- [24] K Pitsilis Georgios, Heri Ramampiaro, and Helge Langseth. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*, 2018.
- [25] Thomas Davidson, Dana Warmusley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515. AAAI Press, 2017.
- [26] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *arXiv preprint arXiv:1802.00393*, 2018.
- [27] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *IcwsM*, pages 61–70, 2015.
- [28] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. *arXiv preprint arXiv:1803.08977*, 2018.
- [29] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *CHI*, pages 3175–3187. ACM, 2017.

- [30] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):174:1–174:30, 2019.
- [31] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *ICWSM*. The AAAI Press, 2014.
- [32] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L Pappa. The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 99–103, 2016.
- [33] Maeve Duggan. Online harassment 2017, 2017. URL <https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017>.
- [34] Aaron Smith and Maeve Duggan. Crossing the line: What counts as online harassment?, 2018. URL <https://www.pewresearch.org/internet/2018/01/04/crossing-the-line-what-counts-as-online-harassment>.
- [35] Brian Dean. Reddit usage and growth statistics: How many people use reddit in 2021?, 2021. URL <https://www.reddithelp.com/en/categories/advertising/managing-ads/engage-comments-thread>.
- [36] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet’s hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32, 2018.
- [37] Shagun Jhaver, Amy S. Bruckman, and Eric Gilbert. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):150:1–150:27, 2019.

- [38] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- [39] Rodrigo Zamith and Seth C Lewis. Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, 659(1):307–318, 2015.
- [40] Fei-Yue Wang, Kathleen M Carley, Daniel Zeng, and Wenji Mao. Social computing: From social informatics to social intelligence. *IEEE Intelligent systems*, 22(2):79–83, 2007.
- [41] Online Harassment Field Manual. Defining “online harassment”: A glossary of terms, 2020. URL <https://onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms>. Accessed: 31 August 2020.
- [42] Chris Benderev. Police arrest man in fatal ‘swatting’ prank, 2020. URL <https://www.npr.org/sections/thetwo-way/2017/12/30/574789231/police-arrest-suspect-in-fatal-swatting-prank>. Accessed: 31 August 2020.
- [43] Ashley Feinberg. The birth of the internet troll, 2014. URL <https://gizmodo.com/the-first-internet-troll-1652485292>. Accessed: 31 August 2020.
- [44] Maja Golf-Papez and Ekant Veer. Don’t feed the trolling: rethinking how online trolling is being defined and combated. *Journal of Marketing Management*, 33(15-16):1336–1354, 2017.
- [45] Judith S Donath et al. Identity and deception in the virtual community. *Communities in cyberspace*, 1996:29–59, 1999.

- [46] Claire Hardaker. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of politeness research*, 6(2):215–242, 2010.
- [47] Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, pages 857–866. International World Wide Web Conferences Steering Committee, 2017.
- [48] Lynette K Watts, Jessyca Wagner, Benito Velasquez, and Phyllis I Behrens. Cyberbullying in higher education: A literature review. *Computers in Human Behavior*, 69:268–274, 2017.
- [49] Janis Wolak, Kimberly J Mitchell, and David Finkelhor. Does online harassment constitute bullying? an exploration of online harassment by known peers and online-only contacts. *Journal of adolescent health*, 41(6):S51–S58, 2007.
- [50] Brian H Spitzberg and Gregory Hoobler. Cyberstalking and the technologies of interpersonal terrorism. *New media & society*, 4(1):71–92, 2002.
- [51] John Bahadur Lamb. Death by swat: The three elements of swatting. In *Video Games Crime and Next-Gen Deviance*. Emerald Publishing Limited, 2020.
- [52] David M Douglas. Doxing: a conceptual analysis. *Ethics and information technology*, 18(3):199–210, 2016.
- [53] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proceedings of the 2017 Internet Measurement Conference*, pages 432–444, 2017.
- [54] Danielle Keats Citron and Mary Anne Franks. Criminalizing revenge porn. *Wake Forest L. Rev.*, 49:345, 2014.
- [55] Mudasir Kamal and William J Newman. Revenge pornography: Mental health implications and related legislation. *Journal of the American Academy of Psychiatry and the Law Online*, 44(3):359–367, 2016.

- [56] Alex Pham. Online bullies give grief to gamers, 2002. URL <https://www.latimes.com/archives/la-xpm-2002-sep-02-fi-grief2-story.html>. Accessed: 31 August 2020.
- [57] Leigh Achternbosch, Charlynn Miller, Christopher Turville, and Peter Vamplew. Grievers versus the grieved—what motivates them to play massively multiplayer online role-playing games? *The Computer Games Journal*, 3(1):5–18, 2014.
- [58] Justin W Patchin and Sameer Hinduja. Cyberbullying and self-esteem. *Journal of school health*, 80(12):614–621, 2010.
- [59] Bonnie S Fisher, Francis T Cullen, and Michael G Turner. Being pursued: Stalking victimization in a national study of college women. *Criminology & Public Policy*, 1(2):257–308, 2002.
- [60] Brianna Wu. Doxxed: Impact of online threats on women including private details being exposed and “swatting”. plus greg lukianoff on balancing offence and free speech. *Index on Censorship*, 44(3):46–49, 2015.
- [61] Nicolas Suzor, Bryony Seignior, and Jennifer Singleton. Non-consensual porn and the responsibilities of online intermediaries. *Melb. UL Rev.*, 40:1057, 2016.
- [62] Ann Pietrangelo. What is verbal abuse? how to recognize abusive behavior and what to do next, 2019. URL <https://www.healthline.com/health/mental-health/what-is-verbal-abuse>. Accessed: 31 August 2020.
- [63] Claude M Steele. Name-calling and compliance. *Journal of Personality and Social Psychology*, 31(2):361, 1975.
- [64] Peter Blau. *Exchange and power in social life*. Routledge, 2017.
- [65] Lawrence E Cohen and Marcus Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608, 1979.
- [66] Philip Purpura. *Security and loss prevention: An introduction*. Butterworth-Heinemann, 2007.

- [67] Sally M Gainsbury, Matthew Browne, and Matthew Rockloff. Identifying risky internet use: Associating negative online experience with specific online behaviours. *New Media & Society*, 21(6):1232–1252, 2019.
- [68] Laura Sofield and Susan W Salmond. Workplace violence: A focus on verbal abuse and intent to leave the organization. *Orthopaedic Nursing*, 22(4):274–283, 2003.
- [69] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31, 2017.
- [70] Jianqing Chen, Hong Xu, and Andrew B Whinston. Moderated online communities and quality of user-generated content. *Journal of Management Information Systems*, 28(2):237–268, 2011.
- [71] Kay Kyeongju Seo. Utilizing peer moderating in online discussions: Addressing the controversy between teacher moderation and nonmoderation. *The American Journal of Distance Education*, 21(1):21–36, 2007.
- [72] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. *arXiv preprint arXiv:1405.1429*, 2014.
- [73] Scott Wright. The role of the moderator: Problems and possibilities for government-run online discussion forums. *Online deliberation: Design, research, and practice*, pages 233–242, 2009.
- [74] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1101–1110, 2008.
- [75] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 839–848, 2013.

- [76] Boreum Choi, Kira Alexander, Robert E Kraut, and John M Levine. Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 107–116, 2010.
- [77] Aiden R. McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Ann Craneheld. Controlling bad behavior in online communities: An examination of moderation work. In *ICIS*. Association for Information Systems, 2016.
- [78] James Grimmelman. The virtues of moderation. *Yale JL & Tech.*, 17:42, 2015.
- [79] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11): 4366–4383, 2018.
- [80] Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 2016.
- [81] Cliff Lampe and Paul Resnick. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550, 2004.
- [82] Cliff Lampe and Erik Johnston. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 11–20, 2005.
- [83] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. Who did what: Editor role identification in wikipedia. In *Tenth International AAI Conference on Web and Social Media*, 2016.
- [84] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc A. Smith. Finding social roles in wikipedia. In *iConference*, pages 122–129. ACM, 2011.
- [85] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):31, 2019.

- [86] Miriam Delgado-Verde, Sarah Cooper, and Gregorio Martín-de Castro. The moderating role of social networks within the radical innovation process: a multidimensionality of human capital-based analysis. *International Journal of Technology Management*, 69(2):117–138, 2015.
- [87] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer, 2017.
- [88] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.
- [89] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.
- [90] Robert J. Topinka. Politically incorrect participatory media: Racist nationalism on r/imgoingtohellforthis. *New Media Soc.*, 20(5):2050–2069, 2018.
- [91] Hsuan-Ting Chen. Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors. *New Media & Society*, 20(10):3917–3936, 2018.
- [92] Brett Sherrick and Jennifer Hoewe. The effect of explicit online comment moderation on three spiral of silence outcomes. *New media & society*, 20(2):453–474, 2018.
- [93] Alex Pham. Reddit moderation tools, 2020. URL <https://mods.reddithelp.com>. Accessed: 31 August 2020.
- [94] Qinlan Shen, Michael Miller Yoder, Yohan Jo, and Carolyn P Rose. Perceptions of censorship and moderation bias in political debate forums. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- [95] Tim Squirrell. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society*, 21(9):1910–1927, 2019.

- [96] Junghee Lee and Hyunjoo Lee. The computer-mediated communication network: exploring the linkage between the online community and social capital. *New Media Soc.*, 12(5):711–727, 2010.
- [97] Seeta Peña Gangadharan. The downside of digital inclusion: Expectations and experiences of privacy and surveillance among marginal internet users. *New Media Soc.*, 19(4):597–615, 2017.
- [98] Marta Cantijoch, Silvia Galandini, and Rachel Gibson. 'it's not about me, it's about my community': A mixed-method study of civic websites and community efficacy. *New Media Soc.*, 18(9):1896–1915, 2016.
- [99] Erin M. Sumner, Luisa Ruge-Jones, and Davis Alcorn. A functional approach to the facebook like button: An exploration of meaning, interpersonal functionality, and potential alternative response buttons. *New Media Soc.*, 20(4):1451–1469, 2018.
- [100] Massimo Ragnedda, Maria Laura Ruiu, and Felice Addeo. Measuring digital capital: An empirical investigation. *New Media Soc.*, 22(5), 2020.
- [101] Peter G Schrader, Mark C Carroll, Michael P McCreery, and Danielle L Head. Mixed methods for human–computer interactions research: An iterative study using reddit and social media. *Journal of Educational Computing Research*, 58(4): 818–841, 2020.
- [102] Mark Cenite, Benjamin H. Detenber, Andy W. K. Koh, Alvin L. H. Lim, and Ng Ee Soon. Doing the right thing online: a survey of bloggers' ethical beliefs and practices. *New Media Soc.*, 11(4):575–597, 2009.
- [103] Nathaniel Poor. Computer game modders' motivations and sense of community: A mixed-methods approach. *New Media Soc.*, 16(8):1249–1267, 2014.
- [104] Chenhao Tan and Lillian Lee. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1056–1066, 2015.

- [105] Natalie Jomini Stroud, Ashley Muddiman, and Joshua M Scacco. Like, recommend, or respect? altering political behavior in news comment sections. *New media & society*, 19(11):1727–1743, 2017.
- [106] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 933–943. International World Wide Web Conferences Steering Committee, 2018.
- [107] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168, 2008.
- [108] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230. ACM, 2017.
- [109] Jonathan Bishop. Representations of ‘trolls’ in mass media communication: a review of media-texts and moral panics relating to ‘internet trolling’. *International Journal of Web Based Communities*, 10(1):7–24, 2014.
- [110] Beatrice Tucker. Student evaluation surveys: anonymous comments that offend or are unprofessional. *Higher Education*, 68(3):347–358, 2014.
- [111] Erik W Black, Kelsey Mezzina, and Lindsay A Thompson. Anonymous social media—understanding the content and context of yik yak. *Computers in Human Behavior*, 57:17–22, 2016.
- [112] Denise Anthony, Sean W Smith, and Timothy Williamson. Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia. *Rationality and Society*, 21(3):283–306, 2009.
- [113] Ruogu Kang, Laura Dabbish, and Katherine Sutton. Strangers on your phone: Why people use anonymous communication applications. In *Proceedings of the*

- 19th ACM conference on computer-supported cooperative work & social computing*, pages 359–370, 2016.
- [114] Wyl McCully, Cliff Lampe, Chandan Sarkar, Alcides Velasquez, and Akshaya Sreevinasan. Online and offline interactions in online communities. In *Proceedings of the 7th international symposium on wikis and open collaboration*, pages 39–48, 2011.
- [115] Kipling D Williams, Cassandra L Govan, Vanessa Croker, Daniel Tynan, Maggie Cruickshank, and Albert Lam. Investigations into differences between social- and cyberostracism. *Group dynamics: Theory, research, and practice*, 6(1):65, 2002.
- [116] Kelly B Filipkowski and Joshua M Smyth. Plugged in but not connected: Individuals’ views of and responses to online and in-person ostracism. *Computers in Human Behavior*, 28(4):1241–1253, 2012.
- [117] Michelle Cleary and Garry Walter. Is e-mail communication a feasible method to interview young people with mental health problems? *Journal of Child and Adolescent Psychiatric Nursing*, 24(3):150–152, 2011.
- [118] Deanna Marie Mason and Bette Ide. Adapting qualitative research strategies to technology savvy adolescents. *Nurse Researcher*, 21(5), 2014.
- [119] José A Casas, Rosario Del Rey, and Rosario Ortega-Ruiz. Bullying and cyberbullying: Convergent and divergent predictor variables. *Computers in Human Behavior*, 29(3):580–587, 2013.
- [120] Howard Rheingold. *The virtual community: Homesteading on the electronic frontier*. MIT press, 2000.
- [121] Ai-Ju Huang, Hao-Chuan Wang, and Chien Wen Yuan. De-virtualizing social events: understanding the gap between online and offline participation for event invitations. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 436–448, 2014.
- [122] Robert B Cialdini and Melanie R Trost. Social influence: Social norms, conformity and compliance. *The handbook of social psychology*, 1998.

- [123] Harry Charalambos Triandis. *Culture and social behavior*. McGraw-Hill New York, 1994.
- [124] Fritz Heider. Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112, 1946.
- [125] Hideyuki Nakanishi, Satoshi Nakazawa, Toru Ishida, Katsuya Takanashi, and Katherine Isbister. Can software agents influence human relations? balance theory in agent-mediated communities. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 717–724, 2003.
- [126] Oliver Posegga and Andreas Jungherr. Characterizing political talk on twitter: A comparison between public agenda, media agendas, and the twitter agenda with regard to topics and dynamics. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.
- [127] Zeynep Tufekci. Can you see me now? audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1):20–36, 2008.
- [128] Suvi Uski and Airi Lampinen. Social norms and self-presentation on social network sites: Profile work in action. *New Media Soc.*, 18(3):447–464, 2016.
- [129] Christina Salmivalli and Marinus Voeten. Connections between attitudes, group norms, and behaviour in bullying situations. *International Journal of Behavioral Development*, 28(3):246–258, 2004.
- [130] Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, 58(6):1015, 1990.
- [131] Icek Ajzen et al. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.
- [132] Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621, 2004.

- [133] Noah J Goldstein, Robert B Cialdini, and Vldas Griskevicius. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, 35(3):472–482, 2008.
- [134] Huifeng Tang, Songbo Tan, and Xueqi Cheng. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760–10773, 2009.
- [135] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.
- [136] Farhan Hassan Khan, Saba Bashir, and Usman Qamar. Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision support systems*, 57:245–257, 2014.
- [137] Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. Using crowdsourcing to improve profanity detection. In *2012 AAAI Spring Symposium Series*, 2012.
- [138] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26, 2012.
- [139] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 105–114, 2019.
- [140] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, volume WS-11-02 of *AAAI Workshops*. AAAI, 2011.
- [141] Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive language and hate speech detection for danish. *arXiv preprint arXiv:1908.04531*, 2019.
- [142] Valentina Sintsova and Pearl Pu. Dystemo: Distant supervision method for multi-category emotion recognition in tweets. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1–22, 2016.

- [143] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.
- [144] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.
- [145] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurovsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118, 2017.
- [146] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, pages 1–6, 2016.
- [147] Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179, 2014.
- [148] Richard Segal, Jason Crawford, Jeffrey O Kephart, and Barry Leiba. Spamguru: An enterprise anti-spam filtering system. In *CEAS*, 2004.
- [149] L Pelletier, Jalal Almhana, and Vartan Choulakian. Adaptive filtering of spam. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 218–224. IEEE, 2004.
- [150] Minoru Sasaki and Hiroyuki Shinnou. Spam detection using text clustering. In *2005 International Conference on Cyberworlds (CW'05)*, pages 4–pp. IEEE, 2005.
- [151] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*, pages 1–10, 2017.
- [152] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*, 2013.

- [153] Politeness. Detecting politeness features in text, 2020. URL <https://cran.r-project.org/web/packages/politeness/vignettes/politeness.html>.
- [154] Miriam A Locher. *Power and politeness in action: Disagreements in oral communication*, volume 12. Walter de Gruyter, 2010.
- [155] TextBlob. Textblob: Simplified text processing, 2020. URL <https://textblob.readthedocs.io/en/dev>.
- [156] Jenny Preece and Jenny Preece. Online communities: Designing usability supporting sociability. *Industrial Management and Data Systems*, 2000.
- [157] Jan-Willem Strijbos and Maarten F De Laat. Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in human behavior*, 26(4):495–505, 2010.
- [158] Scott A Golder and Judith Donath. Social roles in electronic communities. *Internet Research*, 5(1):19–22, 2004.
- [159] Jeffrey Chan, Conor Hayes, and Elizabeth M Daly. Decomposing discussion forums and boards using user roles. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.
- [160] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *WebSci*, pages 13–22. ACM, 2017.
- [161] Anne Mari Undheim and Anne Mari Sund. Prevalence of bullying and aggressive behavior and their relationship to mental health problems among 12-to 15-year-old norwegian adolescents. *European child & adolescent psychiatry*, 19(11):803–811, 2010.
- [162] Estefanía Lozano, Jorge Cedeño, Galo Castillo, Fabricio Layedra, Henry Lasso, and Carmen Vaca. Requiem for online harassers: Identifying racism from political tweets. In *2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 154–160. IEEE, 2017.

- [163] Matthijs Meire, Michel Ballings, and Dirk Van den Poel. The added value of auxiliary data in sentiment analysis of facebook posts. *Decision Support Systems*, 89:98–112, 2016.
- [164] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 133–142, 2011.
- [165] Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linarès. Detection of abusive messages in an on-line community. In *CORIA*, pages 153–168. ARIA, 2017.
- [166] Geoffrey N Leech. *Principles of pragmatics*. Routledge, 2016.
- [167] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.
- [168] Lee Anna Clark, Bruce Cuthbert, Roberto Lewis-Fernández, William E Narrow, and Geoffrey M Reed. Three approaches to understanding and classifying mental disorder: Icd-11, dsm-5, and the national institute of mental health’s research domain criteria (rdoc). *Psychological Science in the Public Interest*, 18(2):72–145, 2017.
- [169] Nicole E Hurt, Gregory S Moss, Christen L Bradley, Lincoln R Larson, Matthew Lovelace, Luanna B Prevost, Nancy Riley, Denise Domizi, and Melinda S Camus. The " facebook" effect: College students’ perceptions of online discussions in the age of social networking. *International Journal for the Scholarship of Teaching and Learning*, 6(2):n2, 2012.
- [170] Abdulwhab Alkharashi, Tim Storer, Joemon Jose, Andrew Hoskins, and Catherine Happer. Understanding abusive behaviour between online and offline group discussions. In *CHI 2019: 37th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI ’19, New York, NY, USA, January 2019. ACM.
- [171] Robyn Caplan. Content or context moderation? artisanal, community-reliant, and industrial approaches. *Data & Society*, November 2018.

- [172] Stephen Connolly, Valentina Klenowski, and Claire Maree Wyatt-Smith. Moderation and consistency of teacher judgement: teachers' views. *British Educational Research Journal*, 38(4):593–614, 2012.
- [173] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. Detecting and tracking political abuse in social media. In *Fifth international AAAI conference on weblogs and social media*, 2011.
- [174] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 91–100. ACM, 2017.
- [175] Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, and Yi Chang. Understanding and discovering deliberate self-harm content in social media. In *Proceedings of the 26th International Conference on World Wide Web*, pages 93–102. International World Wide Web Conferences Steering Committee, 2017.
- [176] Michael Barthel and Amy Mitchell Galen Stocking, Jesse Holcomb. Reddit news users more likely to be male, young and digital in their news preferences. *Pew Research Center*, 2016.
- [177] Jane Wakefield. Anger as shooter video spreads around world, 2019. URL <https://www.bbc.co.uk/news/technology-47583393>.
- [178] UK Home Office. Online abuse and bullying prevention guide, 2015. URL <https://www.gov.uk/government/publications/online-abuse-and-bullying-prevention-guide>.
- [179] Timo Tapani Ojanen, Pimpawun Boonmongkon, Ronnapoom Samakkeekarom, Nattharat Samoh, Mudjaln Cholratana, and Thomas Ebanan Guadamuz. Connections between online harassment and offline violence among youth in central thailand. *Child abuse & neglect*, 44:159–169, 2015.

- [180] Ina Blau and Azy Barak. Synchronous online discussions: Participation in a group audio conferencing and textual chat as affected by communicator's personality characteristics and discussion topics. In *CSEDU (1)*, pages 19–24. INSTICC Press, 2009.
- [181] Marina Bastawrous Wasilewski, Fiona Webster, Jennifer N. Stinson, and Jill I. Cameron. Adult children caregivers' experiences with online and in-person peer support. *Comput. Hum. Behav.*, 65:14–22, 2016.
- [182] David A. Cole, Elizabeth A. Nick, Rachel L. Zelkowitz, Kathryn M. Roeder, and Tawny Spinelli. Online social support for young people: Does it recapitulate in-person social support; can it help? *Comput. Hum. Behav.*, 68:456–464, 2017.
- [183] Jennifer D. Shapka, José F. Domene, Shereen Khan, and Leigh Mijin Yang. Online versus in-person interviews with adolescents: An exploration of data equivalence. *Comput. Hum. Behav.*, 58:361–367, 2016.
- [184] Adrienne Holz Ivory, Jesse Fox, T. Franklin Waddell, and James D. Ivory. Sex role stereotyping is hard to kill: A field experiment measuring social responses to user characteristics and behavior in an online multiplayer first-person shooter game. *Comput. Hum. Behav.*, 35:148–156, 2014.
- [185] Meredith Conroy, Jessica T. Feezell, and Mario Guerrero. Facebook and political engagement: A study of online political group membership and offline political engagement. *Comput. Hum. Behav.*, 28(5):1535–1546, 2012.
- [186] Seong-Jae Min. Online vs. face-to-face deliberation: Effects on civic engagement. *Journal of Computer-Mediated Communication*, 12(4):1369–1387, 2007.
- [187] Sheena L Erete. Engaging around neighborhood issues: How online communication affects offline behavior. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1590–1601. ACM, 2015.
- [188] Hanneke Hendriks, Gert-Jan de Bruijn, Orla Meehan, and Bas van den Putte. Online and offline conversations about alcohol: Comparing the effects of famil-

- iar and unfamiliar discussion partners. *Journal of health communication*, 21(7): 734–742, 2016.
- [189] Sara Vissers and Dietlind Stolle. The internet and new modes of political participation: online versus offline participation. *Information, Communication & Society*, 17(8):937–955, 2014.
- [190] Ella Taylor-Smith and Colin F Smith. Investigating the online and offline contexts of day-to-day democracy as participation spaces. *Information, Communication & Society*, pages 1–18, 2018.
- [191] Matthieu Tixier and Myriam Lewkowicz. Counting on the group: reconciling online and offline social support among older informal caregivers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3545–3558. ACM, 2016.
- [192] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1:1–22, 2017.
- [193] Mainack Mondal, Leandro Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *Proceedings of The 28th ACM Conference on Hypertext and Social Media*, pages 85–94, 2017.
- [194] Dov Cohen. Culture, social organization, and patterns of violence. *Journal of personality and social psychology*, 75(2):408, 1998.
- [195] Donelson R Forsyth. *Group dynamics*. Cengage Learning, 2018.
- [196] Catherine Friend and Nicola Fox Hamilton. Deception detection: The relationship of levels of trust and perspective taking in real-time online and offline communication environments. *Cyberpsychology Behav. Soc. Netw.*, 19(9):532–537, 2016.
- [197] Bassey Etim. Approve or reject: Can you moderate five new york times comments?, 2016. URL <https://www.>

- nytimes.com/interactive/2016/09/20/insider/approve-or-reject-moderation-quiz.html.
- [198] Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linares. Impact of content features for automatic online abuse detection. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 404–419. Springer, 2017.
- [199] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3213–3226. ACM, 2017.
- [200] Paul Graham. How to disagree, 2008. URL <http://paulgraham.com/disagree.html>.
- [201] Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.
- [202] Colleen E Mills, Joshua D Freilich, Steven M Chermak, Thomas J Holt, and Gary LaFree. Social learning and social control in the off-and online pathways to hate crime and terrorist violence. *Studies in Conflict & Terrorism*, pages 1–29, 2019.
- [203] House of Commons. Online abuse and the experience of disabled people, 2019. URL <https://publications.parliament.uk/pa/cm201719/cmselect/cmcompetitions/759/759.pdf>.
- [204] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.
- [205] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.
- [206] Piotr Szymanski and Tomasz Kajdanowicz. Scikit-multilearn: a scikit-based python environment for performing multi-label classification. *The Journal of Machine Learning Research*, 20(1):209–230, 2019.

- [207] Sorin Matei and Sandra J Ball-Rokeach. Real and virtual social ties: Connections in the everyday lives of seven ethnic neighborhoods. *American Behavioral Scientist*, 45(3):550–564, 2001.
- [208] Patrick J Fahy. Use of linguistic qualifiers and intensifiers in a computer conference. *The American Journal of Distance Education*, 16(1):5–22, 2002.
- [209] Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory., 2011.
- [210] Reddit. List of downvoted comments, 2019. URL <https://www.reddit.com/r/ListOfComments/wiki/downvoted>.
- [211] Ashley Muddiman and Natalie Jomini Stroud. News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication*, 67(4): 586–609, 2017.
- [212] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- [213] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heart-mob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):24, 2017.
- [214] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 579–583. IEEE, 2013.
- [215] Tim Weninger, Thomas James Johnston, and Maria Glenski. Random voting effects in social-digital spaces: A case study of reddit post submissions. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 293–297. ACM, 2015.
- [216] Ian Walkinshaw. Agreement and disagreement, 04 2015.

- [217] Jason J Teven, James C McCroskey, and Virginia P Richmond. Measurement of tolerance for disagreement. *Communication Research Reports*, 15(2):209–217, 1998.
- [218] Bryan Frances and Jonathan Matheson. Disagreement. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition, 2018.
- [219] Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013), 2013.
- [220] Lisa Kathryn Hartley, Craig McGarty, and Ngaire Donaghue. Understanding disagreement within the majority about action to atone for past wrongs. *Journal of Applied Social Psychology*, 43:E246–E261, 2013.
- [221] Paul Graham. How to disagree, 2008. URL <http://www.paulgraham.com/disagree.html>.
- [222] Melanie A Revilla, Willem E Saris, and Jon A Krosnick. Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, 43(1):73–97, 2014.
- [223] Amita Misra and Marilyn A. Walker. Topic independent identification of agreement and disagreement in social media dialogue. *CoRR*, abs/1709.00661, 2017.
- [224] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *HLT-NAACL. The Association for Computational Linguistics*, 2003.
- [225] Sara Rosenthal and Kathy McKeown. I couldn’t agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *SIGDIAL Conference*, pages 168–177. The Association for Computer Linguistics, 2015.
- [226] Jie Yin, Nalin Narang, Paul Thomas, and Cécile Paris. Unifying local and global agreement and disagreement classification in online debates. In *WASSA@ACL*, pages 61–69. The Association for Computer Linguistics, 2012.

- [227] Mathieu d'Aquin. Formally measuring agreement and disagreement in ontologies. In *K-CAP*, pages 145–152. ACM, 2009.
- [228] Lu Wang and Claire Cardie. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. *arXiv preprint arXiv:1606.05706*, 2016.
- [229] Fabio Celli, Evgeny Stepanov, Massimo Poesio, and Giuseppe Riccardi. Predicting brexit: Classifying agreement is better than sentiment and pollsters. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 110–118, 2016.
- [230] Reddit. Frequently asked questions, 2019. URL <https://www.reddit.com/wiki/faq>.
- [231] Ben Friedland. Profanity, 2018. URL <https://pypi.org/project/profanity/>.
- [232] Shivam Bansal and Chaitanya Aggarwal. Textstat, 2019. URL <https://pypi.org/project/textstat>.
- [233] Klaus Krippendorff. Reliability in content analysis. *Human communication research*, 30(3):411–433, 2004.
- [234] Justin Cheng, Jaime Teevan, and Michael S Bernstein. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1365–1374. ACM, 2015.
- [235] Chenxi Qiu, Anna Squicciarini, Dev Rishi Khare, Barbara Carminati, and James Caverlee. Crowdeval: A cost-efficient strategy to evaluate crowdsourced worker's reliability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1486–1494. International Foundation for Autonomous Agents and Multiagent Systems, 2018.
- [236] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

- [237] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.
- [238] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *WWW (Companion Volume)*, pages 1100–1105. ACM, 2019.
- [239] Ivan Habernal and Benno Stein Henning Wachsmuth, Iryna Gurevych. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:386–396, 2018.
- [240] Mike Godwin. Godwin’s law of nazi analogies (and corollaries). *EFF.org-Electronic Frontier Foundation*, [WWW document] URL http://w2.eff.org/Net_culture/Folklore/Humor/godwins.law (visited 23.07. 2020), 1995.
- [241] Md Zahidul Islam, Jixue Liu, Jiuyong Li, Lin Liu, and Wei Kang. A semantics aware random forest for text classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1061–1070, 2019.
- [242] Jacob A. Long. *interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions*, 2019. URL <https://cran.r-project.org/package=interactions>. R package version 1.1.0.
- [243] Motherboard. How reddit got huge: Tons of fake accounts, 2012. URL https://motherboard.vice.com/en_us/article/z4444w/how-reddit-got-huge-tons-of-fake-accounts--2.
- [244] Eric Gilbert. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 803–808. ACM, 2013.
- [245] Sarah Talboom and Jo Pierson. Understanding trust within online discussion boards: trust formation in the absence of reputation systems. In *IFIP International Conference on Trust Management*, pages 83–99. Springer, 2013.

- [246] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 2016.
- [247] Jeffrey M Rzeszotarski and Aniket Kittur. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 13–22, 2011.
- [248] Justin Cheng, Moira Burke, and Elena Goetz Davis. Understanding perceptions of problematic facebook use: When people experience negative life impact and a lack of control. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.
- [249] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. "did you suspect the post would be removed?" understanding user reactions to content removals on reddit. *Proceedings of the ACM on human-computer interaction*, 3 (CSCW):1–33, 2019.
- [250] Wilco W Van Dijk, Guido M van Koningsbruggen, Jaap W Ouwerkerk, and Yoka M Wesseling. Self-esteem, self-affirmation, and schadenfreude. *Emotion*, 11(6):1445, 2011.
- [251] Norman T Feather and Rebecca Sherman. Envy, resentment, schadenfreude, and sympathy: Reactions to deserved and undeserved achievement and subsequent failure. *Personality and Social Psychology Bulletin*, 28(7):953–961, 2002.
- [252] Gordon Graham. *The Internet: a philosophical inquiry*. Psychology Press, 1999.

Appendix A

An Appendix

A.1 Pre-discussion questionnaire

1. What is your age?
2. What is your gender?
3. Please state your country of origin.
4. Which of the following is the highest level of education you completed?
5. Which topic category might be an interest of yours? (Check all that apply)
6. Which of the following online social/discussion platforms do you currently have an account with? (Check all that apply)
7. In a typical day, about how much time do you spend using online social/discussion platforms?
8. About how many of your "friends" on social/discussion websites have you met in person?
9. How long have you been using social networking sites?
10. Are you a member of a topic focused group on a social networking communities ? (e.g, football, politics, local community news etc.)

11. Do you accept strangers who try to friend you in social networking sites?
12. How often do you see posts on social media that you think are unacceptable and should be removed or edited?
13. Have you ever been forced to delete one of your own posts?
14. Have you ever been a victim of online abusing (offensive comments or emails)
15. What day do you prefer to join the topic-focused group discussion? (Check all that apply)
16. What time do you prefer to join the topic-focused group discussion? (Check all that apply)
17. Please provide your email address to contact you for further details about the topic-focused group discussion?

A.2 Post-discussion questionnaire

Please read the following instructions before answering the questions:

Read each comment expressed by different groups on both online and in-person settings and use your best judgement to evaluate comments based on the questions and definitions listed below.

To answer the first question, choose one of the following three options:

- A. **Agreeable.** Select this where the comment does not indicate a dispute of some kind.
- B. **Disagreeable.** Select this where the comment embodies counterargument or contradiction.

Counterargument: negation with more supportive arguments.

Contradiction: negation with less or no supportive arguments.

- C. **Abusive.** Select this where the comment embodies ad hominem or name-calling.
Ad hominem: associated with an attack to the character of the person carrying an argument, i.e., a parent who says that the teacher doesn't know how to teach because she graduated from a community college.
Name-calling: abusive language or insults. i.e., *r u stupid!*
- * When you select an answer from the above options, please explain why did you choose your answer.
- ** Choose one of the two following options related to the settings that the comment was made:
- A. **Online:** Refers to a discussion that took a place publicly in the online platform "Group Discussion."
- B. **In-person:** Refers to face-to-face meeting for group discussion in a single place.

A.3 Demographics of participants

Table A.1: Participant corresponding to allocated group and setting. P1-P33 were recruited in in-person group discussion and P34-P67 were recruited in online discussion.

ID	Team	Gender	Age	Education	ID	Team	Gender	Age	Education
P1	T1	M	21-29	High School/GED	P34	T8	M	21-29	High School/GED
P2	T1	M	18-20	High School/GED	P35	T8	M	21-29	High School/GED
P3	T1	F	18-20	High School/GED	P36	T8	M	21-29	Some college
P4	T1	M	18-20	Some College	P37	T8	M	18-20	High School/GED
P5	T1	M	18-20	High School/GED	P38	T8	M	18-20	High School/GED
P6	T2	M	30-39	Bachelor's Degree	P39	T9	M	21-29	Some College
P7	T2	M	18-20	High School/GED	P40	T9	M	21-29	Some College
P8	T2	M	21-29	High School/GED	P41	T9	M	21-29	Bachelor's Degree
P9	T2	M	21-29	Some College	P42	T9	M	21-29	Bachelor's Degree
P10	T3	F	21-29	High School/GED	P43	T9	M	21-29	Bachelor's Degree
P11	T3	F	18-20	High School/GED	P44	T10	M	18-20	High School/GED
P12	T3	M	18-20	High School/GED	P45	T10	F	21-29	High School/GED
P13	T3	F	21-29	High School/GED	P46	T10	M	21-29	High School/GED
P14	T4	F	21-29	High School/GED	P47	T10	F	18-20	Some college
P15	T4	M	21-29	High School/GED	P48	T11	M	18-20	High School/GED
P16	T4	M	21-29	High School/GED	P49	T11	M	18-20	High School/GED
P17	T4	M	21-29	High School/GED	P50	T11	M	21-29	Some College
P18	T4	M	18-20	High School/GED	P51	T11	M	18-20	High School/GED
P19	T5	M	21-29	High School/GED	P52	T11	F	18-20	Some College
P20	T5	M	21-29	Some College	P53	T12	M	18-20	High School/GED
P21	T5	M	21-29	Some College	P54	T12	M	21-29	Some College
P22	T5	M	21-29	High School/GED	P55	T12	F	21-29	High School/GED
P23	T5	M	21-29	Some College	P56	T12	F	18-20	High School/GED
P24	T6	F	21-29	High School/GED	P57	T12	M	18-20	High School/GED
P25	T6	M	21-29	Some College	P58	T13	F	21-29	High School/GED
P26	T6	M	21-29	Some College	P59	T13	M	21-29	High School/GED
P27	T6	M	21-29	High School/GED	P60	T13	M	21-29	High School/GED
P28	T6	F	21-29	High School/GED	P61	T13	M	21-29	High School/GED
P29	T7	M	21-29	High School/GED	P62	T13	M	18-20	High School/GED
P30	T7	F	18-20	High School/GED	P63	T14	M	21-29	Bachelor's Degree
P31	T7	F	18-20	High School/GED	P64	T14	M	18-20	High School/GED
P32	T7	M	18-20	High School/GED	P65	T14	M	21-29	High School/GED
P33	T7	M	21-29	High School/GED	P66	T14	F	21-29	Bachelor's Degree
-	-	-	-	-	P67	T14	F	21-29	High School/GED