A quantitative exploration of repeated advice-seeking on Reddit's r/relationships

Kang, Bianca Ni Ying

2019

Kang, B. N. Y. (2019). A quantitative exploration of repeated advice-seeking on Reddit's r/relationships. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/137027>

<https://doi.org/10.32657/10356/137027>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 19:33:02 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**A QUANTITATIVE EXPLORATION OF REPEATED
ADVICE-SEEKING ON REDDIT'S R/RELATIONSHIPS**

**BIANCA KANG NI YING
SCHOOL OF SOCIAL SCIENCES**

2019

**A QUANTITATIVE EXPLORATION OF REPEATED
ADVICE-SEEKING ON REDDIT'S
R/RELATIONSHIPS**

BIANCA KANG NI YING

School of Social Sciences

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Master of Arts

2019

Statement of Originality

I certify that all work submitted for this thesis is my original work.

I declare that no other person's work has been used without due acknowledgement. Except where it is clearly stated that I have used some of this material elsewhere, this work has not been presented by me for assessment in any other institution or University. I certify that the data collected for this project are authentic and the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

8/11/19

.....
Date

.....
BIANCA KANG

Supervisor Declaration Statement

I have reviewed the content of this thesis and to the best of my knowledge, it does not contain plagiarised materials. The presentation style is also consistent with what is expected of the degree awarded. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

13/11/2019

.....

.....
Assoc Prof Ho Moon-Ho Ringo

Authorship Attribution Statement

This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

8/11/19
.....
Date

.....
BIANCA KANG

Acknowledgements

Foremost gratitude goes to my supervisor Professor Ringo Ho. Thank you for letting me try my hand at this topic though I couldn't even code a "Hello, world!" at the start, and thank you for encouraging me to pursue opportunities whenever I hesitated.

I would also like to offer many thanks to Dr Kokil Jaidka for providing invaluable advice and pointers for the thesis. To the many other academics that I've had the privilege of meeting along the way in spaces online and offline - thanks for hearing and helping a lowly student out.

To my family, thank you for always supporting me. I'm much happier than I would've been chasing my initial contrived fantasies of becoming a clinical practitioner and would never have made it without all of you. To my late grandma - I miss you. To friends, thank you for listening and being there.

To my partner, you were the one who introduced me to this delightful world of text analysis. I think we both know by now that this thesis is just the beginning of a very long dive into that world, and I hope you'll enjoy the journey as much as I have and will.

Lastly, to all the online communities I've ever been a part of and will be a part of. Here's to many more monitor-illuminated late nights and to many more forays into understanding the fascinating nature of being human online.

TABLE OF CONTENTS

Statement of Originality.....	ii
Supervisor Declaration Statement.....	iii
Authorship Attribution Statement.....	iv
Acknowledgements.....	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES.....	x
LIST OF FIGURES.....	xii
ABSTRACT.....	xiii
CHAPTER I: INTRODUCTION.....	1
1.1 Advice and the Internet.....	1
1.2 Significance of thesis.....	2
1.3 Outline of thesis.....	3
CHAPTER II: LITERATURE REVIEW.....	4
2.1 Overview of chapter.....	4
2.2 Advice.....	4
2.2.1 Definition of advice.....	4
2.2.2 Overview of research on advice and advice-seeking.....	5
2.3 Online advice and support communities.....	7
2.3.1 Overview of research on online help-seeking.....	7
2.3.2 Introducing r/relationships.....	9
2.3.3 Distinguishing advice forums from online support groups.....	10
2.3.4 Goal of current study.....	13
2.4 Study setup and variables of interest.....	13
2.4.1 Study setup.....	13
2.4.2 General considerations.....	14
2.4.3 Selected variables and justifications.....	15
CHAPTER III: OVERVIEW OF R/RELATIONSHIPS AND ITS USERS.....	27
3.1 Overview of chapter.....	27
3.2 Purpose of overview.....	27

3.3 Data collection	28
3.3.1 Collection of posts and comments.....	28
3.3.2 Collection of user history.	29
3.4 Topics discussed on the forum	29
3.4.1 Methods.	29
3.4.2 Topics identified.....	30
3.4.3 Summary of results.....	31
3.5 Summarising posting history of r/relationships users	32
3.5.1 Methods.	33
3.5.2 Groups of subreddits identified.	34
3.5.3 Summary of results.....	35
CHAPTER IV: FACTORS RELATED TO REPEAT ADVICE-SEEKING....	36
4.1 Overview of chapter	36
4.2 User selection	36
4.2.1 Removing throwaways.	36
4.2.2 Identifying repeat r/relationships advice-seekers.	36
4.2.3 Identifying non-repeat r/relationships advice-seekers.....	40
4.3 Variables of study	41
4.4 First analysis: repeat vs non-repeat advice-seekers	44
4.4.1 Statistical tests.	44
4.4.2 Prediction task.	46
4.5 Second analysis: repeat vs non-repeat advice-seekers (split by gender)..	49
4.5.1 Justification for creating separate models by gender.....	49
4.5.2 Male user-only prediction model.....	50
4.5.3 Female user-only prediction model.	53
4.5.4 Comments on differences between male/female user models.....	55
4.6 Third analysis: accounting for advice-giving behaviour.....	56
4.6.1 Justification for including advice-giving behaviour.....	56
4.6.2 Creation of user groups.	56
4.6.3 Prediction task.	58
4.7 Fourth analysis: ‘veteran’ repeat advice-seekers vs non-repeat advice-seekers	60

4.7.1 Justification for examining ‘veteran’ repeat advice-seekers.	60
4.7.2 Statistical tests.	60
CHAPTER V: DISCUSSION.....	65
5.1 Overview of chapter.....	65
5.2 Factors related to repeat advice-seeking on r/relationships	66
5.2.1 Demographic and account characteristics.	66
5.2.2 Past activity on certain types of subreddits.	66
5.2.3 More past comments contributed to r/relationships.	67
5.2.4 Content of first r/relationships post.	67
5.2.5 Style of first r/relationships post.....	68
5.3 Going further: considering gender and advice-giving	68
5.4 Comparison with the ‘veteran’ sample	70
5.4.1 Similarities.....	70
5.4.2 Differences.....	70
5.5 Implications.....	72
5.6 Limitations	74
5.7 Future directions.....	75
REFERENCES	78
APPENDIX A: TOPIC MODELLING ON POSTS AND COMMENT GROUPS.....	96
APPENDIX B: CREATION OF THEMATICALLY-SIMILAR GROUPS OF SUBREDDITS.....	102
APPENDIX C: CREATION OF VARIABLES OF STUDY.....	117
APPENDIX D: MODEL TRAINING AND SELECTION FOR FIRST SET OF ANALYSES (REPEAT VS NON-REPEAT ADVICE-SEEKERS).....	120
APPENDIX E: RESULTS OF KRUSKAL-WALLIS TESTS ON REPEAT/NON-REPEAT ADVICE-SEEKERS SPLIT BY GENDER.....	125
APPENDIX F: MODEL TRAINING AND SELECTION FOR SECOND SET OF ANALYSES (MALE REPEAT VS MALE NON-REPEAT ADVICE-SEEKERS).....	142
APPENDIX G: MODEL TRAINING AND SELECTION FOR SECOND SET OF ANALYSES (FEMALE REPEAT VS FEMALE NON-REPEAT ADVICE-SEEKERS)/.....	146

APPENDIX H: MODEL TRAINING AND SELECTION FOR THIRD SET OF ANALYSES (ACCOUNTING FOR ADVICE-GIVERS).....	150
---	-----

LIST OF TABLES

Table 1. <i>149 variables used in current study.</i>	42
Table 2. <i>Significant results after Holm-Bonferroni correction (repeat vs non-repeat advice-seekers).</i>	45
Table 3. <i>Normalised confusion matrix for model from first set of analyses.</i>	47
Table 4. <i>Variables positively/negatively related to repeat advice-seekers.</i>	47
Table 5. <i>Normalised confusion matrix for male-only model from second set of analyses.</i>	51
Table 6. <i>Variables positively/negatively related to male repeat advice-seekers</i>	52
Table 7. <i>Normalised confusion matrix for female-only model from second set of analyses.</i>	54
Table 8. <i>Variables positively/negatively related to female repeat advice-seekers.</i>	54
Table 9. <i>Breakdown of n=14155 user activity in 3 months from first post in dataset.</i>	58
Table 10. <i>Normalised confusion matrix for multinomial model from third set of analyses.</i>	59
Table 11. <i>Significant results after Holm-Bonferroni correction ('veteran' repeat vs non-repeat advice-seekers).</i>	61
Table A1. <i>28 topics retrieved from 70520 posts and corresponding keywords.</i>	97
Table A2. <i>36 topics retrieved from 68925 comment groups and corresponding keywords.</i>	99
Table B1. <i>31 groups of subreddits retrieved by community detection algorithm.</i>	104
Table B2. <i>44 groups of subreddits refined from original Personal/Sharing group.</i>	108
Table D1. <i>Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – base model.</i>	121
Table D2. <i>Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – base model.</i>	122
Table D3. <i>Non-zero coefficients (base model).</i>	123
Table E1. <i>Significant results of Kruskal-Wallis tests after Holm-Bonferroni correction.</i>	126
Table F1. <i>Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – male-only model.</i>	142
Table F2. <i>Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – male-only model.</i>	143
Table F3. <i>Non-zero coefficients (male-only model).</i>	145
Table G1. <i>Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – female-only model.</i>	146

Table G2. <i>Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – female-only model.</i>	147
Table G3. <i>Non-zero coefficients (female-only model).</i>	148
Table H1. <i>Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – multinomial model.</i>	151
Table H2. <i>Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – multinomial model.</i>	152
Table H3. <i>Coefficients from multinomial logistic regression model.</i>	154

LIST OF FIGURES

Figure 1. <i>Screenshot of r/relationships (2018)</i>	2
Figure 2. <i>Flow of analyses covered in Chapter 3.</i>	28
Figure 3. <i>Selecting repeat advice-seekers with no prior r/relationships posts.</i> ..	39
Figure 4. <i>Selecting repeat advice-seekers with prior r/relationships posts.</i>	40
Figure 5. <i>Selecting non-repeat advice-seekers.</i>	41

ABSTRACT

Online spaces for users seeking advice for personal issues have existed since the introduction of the Internet to the public in the 1990s. Today, online advice-seeking is a commonplace activity. However, understanding of this phenomenon is still limited. Much research on advice has focused on the evaluation and outcomes of advice, and on offline exchanges between non-stranger individuals. The current study was thus an exploratory step focused on identifying the factors related to repeated online advice-seeking behaviour. Specifically, the current study examined how (1) user-level differences and (2) differences in forum community reception may be associated with a user deciding to seek advice again on the same advice forum. One years' worth of postings from a popular relational advice forum (Reddit's r/relationships) and the posting activity of its users were analysed with the help of computational methods. Results indicated that certain user-level cues were predictive of repeat advice-seeking behaviour. Further analyses highlighted the importance of accounting for gender even when examining behaviour in a pseudonymous faceless environment. Practical implications and future directions were proposed and discussed.

CHAPTER I: INTRODUCTION

1.1 Advice and the Internet

The Internet has transformed advice-seeking. Where previously advice-seekers were confined to close offline ties or – for those seeking some modicum of anonymity – local aunt agony columns, there is now an abundance of anonymous/pseudonymous advice forums online for almost every type of problem. However, despite how ubiquitous advice-seeking is in today's online landscape, understanding of this phenomenon is still limited. Current research paradigms of advice are largely informed by offline advice studies (MacGeorge, Feng & Guntzviller, 2016a) which rely heavily on controlled settings/surveys studying individuals known to each other. There is also a general dearth of research on the decision to seek advice, even in offline settings (Brooks, Gino & Schweitzer, 2015).

Early research on communities related to online advice forums (e.g. online support groups) have noted how features like anonymity and accessibility attract usage (e.g. Cotten & Gupta, 2004) but do not fully address more nuanced differences in user behaviours. One obvious usage distinction lies between the users who return to the same community to seek for more help versus those who do not. A related line of work on online support groups has investigated user return as a form of user retention and examined how social feedback from other members influences the likelihood of a user staying active within the support community (e.g. Wang, Kraut & Levine, 2015). However, little is known about how user-level differences may also be related to repeated help-seeking, and how all these factors may play out in an online advice forum (which may follow different social dynamics between users; see 2.3.3 Distinguishing advice forums from online support groups).

The main goal of the current study was thus to explore how user-level differences and differences in community reception may be associated with the user returning to seek for advice on the same advice forum.

1.2 Significance of thesis

The thesis provides an initial exploratory look into online advice-seeking in a non-health context – a popular but relatively understudied online behaviour (MacGeorge et al., 2016a; Rains & Wright, 2016). The current study focuses on a large relationship advice forum of over two million subscribers (redditlist.com, 2019), Reddit's r/relationships (see Figure 1 for a screenshot), which has received much popular media attention but very little academic attention. By selecting such a forum for study, the thesis hence also has practical importance beyond its theoretical interests in providing a more nuanced understanding of online advice-seeking. For example, a better understanding of features related to repeat advice-seeking can assist future efforts in early identification of users who are likely to request for the forum's advice more than others. Moderators of such forums can then opt to pay more attention to the comments delivered to such users, especially since repeated advice-seeking may be suggestive of greater reliance or even trust in the forum for guidance.

Figure 1. Screenshot of r/relationships (2018)

More generally, an increasing amount of human activities and interactions are occurring and being recorded online (Lambiotte & Kosinski, 2014). The Internet is a huge repository of databases of human behaviour, far more comprehensive and larger than traditional psychology datasets (Gonzalez-

Bailon, 2013). For the study of specific behaviours like advice-seeking, such datasets also circumvent the social desirability biases that may arise during self-report measures (Barbera, Jost, Nagler, Tucker & Bonneau, 2015) and the limitations of recollection in retrospective surveys. However, technical barriers to collecting and processing such data may have previously put many social scientists off from utilising them (Kosinski, Wang, Lakkaraju & Leskovec, 2016). This study adds to the nascent but fast-growing pool of theory-driven social science research on large datasets of online behaviour.

1.3 Outline of thesis

Chapter 2 reviews relevant work that has been done in the areas of advice and online help groups and introduces r/relationships. Following that, it explicates how the main goal of the study was derived, describes the variables of study, and explains why they were chosen.

As mentioned, the r/relationships forum has not been thoroughly studied formally. Thus, Chapter 3 covers the methods and results of a preliminary exploration undertaken by the current study to better understand the topics of discussion on r/relationships and the posting histories of its advice-seeking users.

Chapter 4 covers the main analyses and results of the study. All analyses were targeted at exploring the features that differentiate users who returned to seek for more advice on r/relationships versus users who did not. Four sets of analyses were conducted – the flow of analyses is detailed in Chapter 2 (specifically, 2.4 Study setup and variables of interest).

Chapter 5 summarises the findings of the thesis and discusses its implications. Limitations of the current study are highlighted before concluding with suggestions for future directions.

CHAPTER II: LITERATURE REVIEW

2.1 Overview of chapter

This chapter reviews concepts core to the study (i.e., advice, online help-seeking, and online help groups) and outlines the main goal of the study. It also covers the variables of study and the justifications for selecting them.

2.2 Advice

2.2.1 Definition of advice. Advice is a universal act of communication; it is sought for and shared between people every day, present across age groups and cultures (MacGeorge et al., 2016a). Generally, advice can be defined as a proposal or suggestion from the advisor regarding ways the advisee may tackle a given problem (Locher & Hoffman, 2006; MacGeorge et al., 2016a). At its crux, advice hence is a subjective prescription for a certain course of action judged by the advisor as appropriate for the advisee (Locher & Hoffman, 2006; Feng, Jang, Kim & Wang, 2017) and involves some degree of persuasion (Feng & Burleson, 2008). The act of advice is often linked to the larger concept of social support (e.g. Smithson et al., 2011). It is frequently categorised under the umbrella of informational support (e.g. Cutrona & Suhr, 1992; Goldsmith, 2004; Ruble, 2011). Informational support can be understood as information given to the support-seeker on how to handle their issues (House, 1981).

Note that the original definition of advice does not ascertain that it is necessarily supportive (MacGeorge et al., 2016b). In fact, multiple studies have indicated that individuals receiving advice may not necessarily view it as beneficial (Goldsmith, 1999). A key reason for why advice may vex instead of help is because of its threat to face (Goldsmith, 1999). Face is a concept introduced by Brown and Levinson (1987) and is related to the public perception that individuals wish to uphold for themselves. Face may be lost through the act of receiving advice – for example, seeking advice may be seen as an admission of one's ineptitude (Sillence, 2010), or the advice may be

delivered in a way that chastens the advice-seeker's feelings of independence (Goldsmith & MacGeorge, 2000).

2.2.2 Overview of research on advice and advice-seeking. Many studies, especially from the communication paradigm, have been conducted on advice. Common research methods involve individuals known to the participant (e.g., friends, romantic partners) (e.g., Goldsmith, McDermott & Alexander, 2000; MacGeorge et al., 2016b), asking participants to read and rate a few researcher-constructed scenarios (e.g. Caplan & Samter, 1999; MacGeorge, Lichtman & Pressey, 2002; Feng & Burleson, 2008; Feng et al., 2017), and filling in questionnaires or rating recalled experiences with advice (e.g. Goldsmith et al., 2000; Feng & MacGeorge, 2006; Feng & MacGeorge, 2010; Carlson, 2016). In the organisational psychology field, advice studies have focused largely on intellectual tasks (i.e., tasks with an objectively right or wrong answer) (e.g. Yaniv 2004; Poston, Akbulut & Looney, 2005; Bonaccio & Dalal, 2006). On a whole, these advice studies follow two main themes – (1) examining the recipient's *response* or evaluation of advice given in offline settings and/or in hypothetical scenarios (e.g. Caplan & Samter, 1999; Goldsmith, 1999; Goldsmith & MacGeorge, 2000; Goldsmith et al., 2000; MacGeorge et al., 2002; Dalal & Bonaccio, 2010; Feng & Burleson, 2008; Feng & MacGeorge, 2010; MacGeorge et al., 2016b), and/or (2) *advice-taking* in offline settings (e.g. Yaniv, 2004; Mansell & Lam, 2006; Van Swol, 2011, Tost, Gino & Larrick, 2012; Yaniv & Choshen-Hillel, 2012; Carlson, 2016; Van Swol, MacGeorge & Prah, 2017). The advice-response theory has also been formulated by MacGeorge and colleagues; it attempts to integrate advisor features (e.g. expertise) with features of the advice given (e.g. politeness) in a bid to understand outcomes of receiving advice (e.g. intention to use the advice) (MacGeorge et al., 2016a).

There is thus firstly a lack of research on the earlier stage of the advice process; specifically, the stage of advice-seeking (Gino, Brooks & Schweitzer, 2012; Brooks et al., 2015). Brooks et al. (2015) remarked that advice-seeking is

in fact usually left out of typical experimental setups for studying advice. The limited knowledge from studies that have explicitly studied advice-seeking was summarised by Brooks et al. (2015) – notably, advice-seekers generally decide to look for advice when they are unsure, anxious, or facing a difficult problem. However even then, these studies – and Brooks et al.’s (2015) – were not contextualised explicitly for online advice settings.

This leads to the second literature gap – a lack of online-specific advice studies, despite the proliferation of very topically diverse question-and-answer sites like Yahoo Answers, Quora, and large advice forums like Reddit’s r/relationships subreddit¹. Again, current research paradigms of advice are still largely informed by offline advice studies (MacGeorge et al., 2016a) and face-to-face scenarios (Rains and Wright, 2016). Many studies of advice have been conducted in the context of intimate and friendly relationships (MacGeorge et al., 2016a). Furthermore, many advice studies have used participants of mostly undergraduate age (MacGeorge et al., 2016a) – so even though studies have attempted to simulate online settings, they have still been limited to scenarios that chiefly pertain to undergraduates (e.g. failing an examination) (e.g. Feng, Li & Li, 2016; Rains, Tsetsi, Akers, Pavlich & Appelbaum, 2018).

Both these gaps are critical to address given that findings from previous studies of offline advice may not translate fully to the online experience. Researchers have suggested that individuals who choose to utilise online relationship advice boards and those who prefer face-to-face advice may be unlike each other (Collison et al., 2018). Furthermore, in an online setting it may no longer be possible to neatly separate concepts like advisor characteristics from the messages that they produce (as done in offline settings with the advice-response theory, e.g. Feng and MacGeorge, 2010). When online – especially in pseudonymous environments with little profile information – one is no longer privy to cues beyond whatever is displayed on the screen.

¹ Reddit is a highly popular social news site; a subreddit can be thought of as a discussion forum on Reddit focused on a specific topic.

Findings like how advice given from persons not close to the individual tend to be seen as improper (Goldsmith & Fitch, 1997) also no longer hold well in the mass online context, where people purposely seek advice from strangers and weak ties. In fact, many online settings appear to follow a completely different set of norms from face-to-face relations. For instance, users typically jump right into the details of their problems and in quite intimate fashion despite a lack of prior interaction history (Walther & Boyd, 2002).

In conclusion, though a vast literature on advice exists, the areas of advice-seeking and online advice remain relatively understudied. The mass availability of large amounts of organic data on online behaviour and interactions (Landers, Brusso, Cavanaugh & Collmus, 2016) offers a promising avenue in which to further explore both areas. It was on this basis that the general premise of the current study was formulated – to explore the factors related to the specific behaviour of repeat advice-seeking on an online advice forum. In other words, the study sought to better understand why certain users chose to seek online advice more often than others and if there were identifiable factors that distinguished the two user groups. The online advice forum selected was Reddit's r/relationships, one of the largest advice forums on Reddit with over two million subscribers (redditlist.com, 2019). The study's premise and justification for choosing r/relationships are further discussed in the following section (2.3 Online advice and support communities).

2.3 Online advice and support communities

2.3.1 Overview of research on online help-seeking. A main focus of early research on online help-seeking (i.e., to seek support and/or information) was the identification of general factors that attract people to take the first step to go online. Online help communities were postulated to have a whole range of appealing features for potential users; amongst others: anonymity, accessibility, absence of stigma and greater openness of discussion of thornier issues, an ability to decide when to connect and disconnect, and a diversity of viewpoints (White & Dorman, 2001; Walther & Boyd, 2002; Colvin, Chenoweth, Bold &

Harding, 2004; Cotten & Gupta, 2004, Suzuki & Calzo, 2004; Christopherson, 2006; Leimester, Schweizer, Leimester & Krcmar, 2008; Coursaris & Liu, 2009). These features likely still hold today; a recent study with Reddit users (Redditors) found that they appreciated the freedom to post openly about difficult issues (Shelton, Lo & Nardi, 2015).

However, these general features of online help communities are not fully able to explain more nuanced differences in user behaviour. An associated line of research has thus examined the differences between people who seek online support and those who do not. McKenna and Bargh (1999) suggested that people who feel unable to express core parts of themselves in their offline relationships, and anxious and lonely people may be more disposed to head online. Personality may also play an important role (Hughes, Rowe, Batey & Lee, 2012; Lo, 2019); neuroticism was positively related to greater use of Facebook for support and agreeableness negatively related (Giota and Kleftras, 2014).

A related and even more specific angle that research can and has taken is the study of how online interactions pan out differently *within* the group of people who seek online help. This domain of research is rapidly expanding given that communication online and available records of such interactions are increasing swiftly (Lambiotte & Kosinski, 2014). Many studies have focused on the *response* to the users from the respective communities. One common area of investigation is the relationship between the response of the community and the content of the initial post written by the help-seeking user. For instance, usage of words related to family and prayer in Facebook posts about surgery was found to be related to more comments (Davis, Anthony & Pauls, 2015). On altruistic communities like Reddit's Random Acts of Pizza, expressions of gratitude and clear conveyance of need in the posts of users seeking a free pizza were among the factors related to their success in getting a pizza from the community (Althoff, Danescu-Niculescu-Mizil & Jurafsky, 2014). Users who

self-disclosed more in their posts received comments more swiftly and in greater quantities on mental health subreddits (Balani & De Choudhury, 2015).

A number of studies have taken a more longitudinal view and examined how the response of the community is in turn related to future user behaviour. A key user behaviour of interest to the current study is the returning of the user to the community to post again for help. In previous studies examining user return (e.g. Arguello et al., 2006; Joyce & Kraut, 2006; Wang, Kraut & Levine, 2012; Wang et al., 2015; Cunha, Weber, Haddadi & Pappa, 2016), user return has typically been conceptualised as a form of user retention and generally viewed as a positive behaviour that the community should aim to upkeep. However, the studied communities were not advice forums; they were online support groups. The distinctions between these support and advice communities (especially r/relationships itself) are important and had implications on the specific framing of the current study. A brief introduction of the r/relationships advice forum is given below, followed by a discussion of how the differences between support and advice communities influenced the scoping of the study.

2.3.2 Introducing r/relationships. r/relationships is an advice forum for issues centred around human relationships. The relationships do not necessarily need to be romantic; any sort of relational issue – from the familial to platonic, and so on – may be sought for advice on the forum (r/relationships wiki, 2019). Redditors submit text posts that have a 40000-character limit (“[reddit change] Selfpost character limit is now 40,000 for all subreddits : changelog”, 2015) detailing a specific relationship problem they would like other Redditors to advise them on. Reddit posts are typically written in Standard English (Schrading, Alm, Ptuchas & Homan, 2015). Other Redditors may then respond to the post by voting and/or commenting with their advice.

Justification for selecting r/relationships. A main reason for selecting r/relationships as the advice forum of study was its clear structure of discussion. Users post about a specific problem and comments replying directly to the post

are focused on providing advice. This provided a clean unambiguous setup for collecting and using the data to understand online advice-seeking.

Furthermore, despite the size of r/relationships and the popular media attention surrounding the forum (e.g. Mulshine, 2014; Ahmed, 2017; Smith, 2018), it has not received the same amount of academic attention as the mental health and suicide support subreddits. So far there has been only one study examining the language of online advice using r/relationships by Collisson et al. (2018), which evaluated a set of 117 posts discussing only romantic relationships. The more general content of r/relationships (i.e., interpersonal relationships) offers a valuable opportunity to expand the literature on support communication which has largely focused on health-related conditions (Ruble, 2011; Rains & Wright, 2016). Advice-seeking online for relationships, especially of the romantic variety, is itself also a ripe area for more research; previous studies have highlighted the pervasiveness of such topics on teenage advice forums (Suzuki & Calzo, 2004; Weinstein & Selman, 2014). In addition, Reddit is relatively understudied in academic research (Olson & Neal, 2015; Dehghani et al., 2017) versus social networking services like Facebook and Twitter.

2.3.3 Distinguishing advice forums from online support groups.

Generally, online support groups and advice forums like r/relationships are similar in that they are both spaces targeted at providing support and help for users who seek it. Both types of online communities are also generally made up of users with little to no prior offline relationships with each other, in contrast to social networking services like Facebook (boyd & Ellison, 2008; Shelton et al., 2015; Tomprou, Dabbish, Kraut & Liu, 2019).

Still, despite the apparent similarities between online support groups and online advice forums, there may be differences in the interaction dynamics within them. Online support groups are usually characterised as a community of users bonded by their facing of similar issues, e.g. mental health problems (Walther & Boyd, 2002; Lyons, Aksayli & Brewer, 2018; Fraga, da Silva &

Murai, 2018; Li, Coduto & Morr, 2019). Users post regular updates to the group (Tichon & Shapiro, 2003; Colvin et al., 2004; Chung & Chen, 2018). In addition, reciprocity of support between users is commonly highlighted as a key feature of these groups (e.g. Joyce & Kraut, 2006; Barak, Boniel-Nissim & Suler, 2008; Hwang et al., 2010). In other words, it is customary for users to act both as regular providers *and* recipients of support. Smithson et al. (2011) in fact observed how users of an online self-harm support group appeared cognizant of this convention and would apologise in their writings if they failed to adhere to it. Given this characterisation of online support groups, it is then unsurprising that user return would be conceptualised as definitively desirable. As explained by Wang et al. (2015), users who consistently participate in such online support groups not only reap the benefits of receiving support, but also act as valuable sources of support to the rest of the community.

On the other hand, the r/relationships advice forum disallows frequent and multiple update posts about a problem (such posts are actually taken down by its moderators) (r/relationships wiki, 2019). Posts are expected to be clear about the problem that the user wishes to seek advice for, not meandering or vague off-my-chest rants (r/relationship wiki, 2019). Such a setting is very different from physical and/or mental health support groups, and even from other mental health support subreddits that have been previously examined (e.g. in De Choudhury & De, 2014; De Choudhury & Kiciman, 2017). Reciprocity norms may also be weaker or lacking, based on the posting patterns observed in studies of other advice forums. In a study of Yahoo Answers, it was noted that for the Marriage category, there was some – but very little – overlap between answer-givers and answer-seekers (Adamic, Zhang, Bakshy & Ackerman, 2008). In a qualitative study of 273 threads on teenage health and sexuality boards spanning three months, the vast majority of users asked only one or two questions in that timeframe, and a great bulk of users also shared only about one to three replies to the questions of others (Suzuki & Calzo, 2004).

Hence in the context of an advice forum, especially one like r/relationships, it is less clear whether user return is definitively positive. While in online support communities, repeated postings from the user may be inferred to be acts of commitment toward the group (Arguello et al., 2006; Joyce & Kraut, 2006) (e.g. postings to show solidarity/encouragement), it is less clear for an advice forum like r/relationships. Repeat posts by a user on r/relationships cannot be fully interpreted as an effort at community-bonding since the only posts that are allowed are expected to be those asking for advice on dealing with a user-specific problem. The current study thus makes no claim on whether user return on r/relationships is positive or negative for the forum and/or the user themselves; concluding either lies beyond the scope of this project. In the current scenario, the more conservative interpretation that can be made is that the user is displaying some degree of reliance on the advice of r/relationships.

Another consideration was regarding the role of the community's response to the user in influencing their decision to return to the forum. While the community's response took centre-stage in the studies on user return in online support groups, it was less apparent how important it would be in the context of a forum like r/relationships. On a weight loss support subreddit, a positive welcome (via the number of positive comments and upvotes²) was found to be related to whether a user continued posting (Cunha et al., 2016). Similarly, receiving a reply was related to users returning to post and/or comment in Arguello et al.'s (2006) study of online support groups. However, since r/relationships does not allow posts outside of those with a specific problem, while the community may respond positively to the initial post, the impact of their response on whether the user will post again for advice may be weaker. Users may simply not return because they do not have a problem that needs advising.

² Upvotes are given by other users to a user's posts and comments. Content is upvoted if it is thought by the voting user to be relevant/contributes to the discussion and is an expression of support for the content.

2.3.4 Goal of current study. The overall goal of the current study was to explore the relationships between individual-level user differences and the decision to seek advice again on the same advice forum. At the same time, the current study recognised the importance of the community given past research and also attempted to account for the variability of its response to different users.

Specifically, the current study explored how: (1) user-level differences and (2) differences in forum community reception may be associated with the decision to seek advice again on the same advice forum. A user was considered to have decided to seek advice again on r/relationships (i.e., be a repeat advice-seeker) if they created a new advice-seeking post within three months after their first post on r/relationships.

2.4 Study setup and variables of interest

2.4.1 Study setup. As stated earlier, the focal goal of the current study was to explore how (1) user-level differences and (2) differences in forum community reception may be associated with a user deciding to seek advice again on the same advice forum. It is important to note that given the observational nature of the data utilised in the study, all analyses describe correlation and do not claim to depict causality (Schwartz & Ungar, 2015).

The study began by getting an exploratory overview of the forum to better understand it. Specifically, the study sought to first understand (1) what was being discussed on r/relationships, using an unsupervised topic modelling algorithm on the collected postings, and (2) where its advice-seeking users have been on Reddit, using network analysis and text clustering to summarise their posting histories. This preliminary exploration is covered in Chapter 3.

The main analyses are covered in Chapter 4. The analyses began with a base comparison to identify the differences between posters who returned to seek for more advice versus posters who did not return to seek for more advice. For the second set of analyses, the posters were split by gender for separate comparisons for male and female users. The important features for predicting a

repeat advice-seeker/non-repeat advice-seeker were noted for each gender and compared. The third set of analyses accounted for the additional behaviour of *advice-giving*. Specifically, it focused on identifying and separating: (1) users who came back to give *and* seek advice, (2) users who came back to seek but not give advice, (3) users who came back to give but not seek advice, and (4) users who did not come back to give or seek advice. Finally, the last set of analyses compared ‘veteran’ repeat advice-seekers (i.e., repeat advice-seekers who had posted for advice on r/relationships prior to their first post in the collected one-year dataset) and non-repeat advice-seekers.

2.4.2 General considerations. The current study worked on one year’s worth of posts and comments from r/relationships and the Reddit activity of the examined r/relationships posters (see 3.3 Data collection, for data collection details). Given the large sample size, it was likely that most statistical tests would prove significant (Adjerid & Kelley, 2018). Thus, in the current study, corrections were applied to control for multiple comparisons (Domingos, 2012) and effect sizes reported (Adjerid & Kelley, 2018). Where possible, predictive models with the selected variables were also trained for testing on a held-out dataset. This was done to ensure that the identified relationships between the variables and the outcome of interest (deciding to seek advice again) were not spurious and generalisable beyond the training set (Adjerid & Kelley, 2018). These are detailed in Chapter 4.

Given the size of the dataset, there were many possible variables to study. A theory-driven approach to the large amount of data (e.g. Chen & Wojcik, 2016; Landers et al., 2016) was used. Such an approach can assist in identifying biases in the data and in interpreting the patterns found in a coherent, theoretically satisfying way (Gonzalez-Bailon, 2013). As much of the collected data was raw text, a core component of the project thus was defining and extracting psychologically relevant features from the written text to represent the user-level and community-level variables. Due to the large amount of data, computational methods instead of qualitative coding were

relied on to quantify features in the text efficiently (Kern et al., 2016). The study of language by computerised text analysis is not new to the field of psychology, dating back to the 1960s (Moessner, Feldhege, Wolf & Bauer, 2018). Many studies by Pennebaker and colleagues have demonstrated that the different ways in which people express themselves can be captured quantitatively and that these quantified variables are related to individual differences; whether in permanent ways like the manner in which individuals see the world and their personalities, or more transient influences like their feelings at the time of writing (Biber, 1988; Pennebaker & King, 1999; Chung & Pennebaker, 2007; Pennebaker, Chung, Frazee, Lavergne & Beaver, 2014; Boyd & Pennebaker, 2017). Chapters 3 and Appendices A to C describe how user-level and community-level feature extraction and creation were automated with the help of computational methods. Besides language, the project also made use of available metadata (e.g. post scores) and user activity history (i.e., past posts and comments on other subreddits).

Overall, the project was focused on the content produced by r/relationships posters instead of connections between the users. This decision reflects how Reddit generally focuses more on user-submitted content in contrast to connection-centric sites like Facebook (Pappa et al., 2017; Fraga et al., 2018). The following section (2.4.3 Selected variables and justifications) lists the textual, metadata, and activity variables involved in the study, along with the relevant justifications for selecting them. The final full list of variables can be found in 4.3 Variables of study.

2.4.3 Selected variables and justifications. This section lists the variables identified for study in the project. The relevant literature backing their selection is described and any expected relationship between the variables and the decision to seek advice again stated.

Demographics: age and gender. Age and gender may be important factors influencing the seeking and receiving of online support. Young adults spend more time connecting with strangers online (Thayer & Ray, 2006). Being

younger and female was associated with pursuing more online interaction and support for health-related issues (DeAndrea & Anthony, 2013; Mattsson, Olsson, Alfonsson, Johansson & Carlsson, 2015).

Age and gender also appear to have non-trivial relations with how individuals express themselves in writing. Older individuals display less focus on themselves and more on others, and make more references to the present and future (Pennebaker & Stone, 2003; Tausczik & Pennebaker, 2010). Male writers typically use more articles (e.g. ‘the’, ‘a’) (Chung & Pennebaker, 2007; Tausczik & Pennebaker, 2010), which researchers have interpreted as reflective of a more specific and concrete thinking style (Chung & Pennebaker, 2007). Female writers on the other hand generally use more verbs, pronouns, and social-related words (e.g. ‘help’, ‘speak’) – researchers have construed this as indicative of the relatively higher social inclinations of women (Chung & Pennebaker, 2007; Tausczik & Pennebaker, 2010). In self-disclosures of depression on Twitter, male users were found to express themselves more angrily than female users, who typically expressed themselves using more anxiety- and sadness-related words (De Choudhury, Sharma, Logar, Eekhout & Nielsen, 2017).

The user’s gender and style of writing together may in turn be crucial in considering the community’s response. An examination of support on forums from StackExchange, Reddit, and Wikipedia found that having a female-gendered username and female manner of writing were related to greater amounts of supportive *and* unsupportive comments (Wang & Jurgens, 2018). Teenagers on a romantic relationship advice forum were observed to provide more suggestions for seeking help and breaking up to female posters versus male posters (Kim, Weinstein & Selman, 2016).

Relevance to current study. Given that younger users appear more open to interacting with strangers online, it was expected that younger age in r/relationships users would be related to returning to seek for more advice. Identifying as female was also expected to be related to returning for advice.

Based on the past literature describing the many ways in which the two genders differ in writing style, it was surmised that these differences would also arise in this study. Contingency was thus also planned for such differences – as mentioned, besides building a base model that included both genders, separate models for each gender were also constructed.

Reddit account age. In studies of other virtual communities – like online travel discussion groups – the amount of time the user had been registered to the site was understandably related to the self-reported intent to continue using the community as a source of information (Zhao, Stylianou & Zheng, 2013). In the context of Reddit, there are observable differences in behaviour between the users based on cohort year – Redditors who joined the site earlier in its history (2009 to 2011) generally comment more than they post (Barbosa, Cosley, Sharma & Cesar-Jr, 2015).

Relevance to current study. It was unclear from the literature if there was a specific direction that could be expected for the relationship between Reddit account age and seeking advice again. Nonetheless, account age was included for analysis as it likely had implications for observed user behaviour.

Activity on Reddit prior to first r/relationships post. The activity history of Redditors may provide valuable information about their future behaviour and interactions. For example, activity in another recovery-related subreddit (r/REDDITORSINRECOVERY) was characteristic of longer-term drinking abstiners who were users of the drinking abstinence community r/StopDrinking (Tamersoy, De Choudhury & Duen, 2015). Prior activity on mental health subreddits was positively linked to posters receiving more comments on r/RoastMe³ (Kasunic & Kaufman, 2018). Elsewhere on Yahoo Answers, it was found that users typically participated within a narrow span of forum categories despite the great variety of topical forums available (Adamic et al., 2008).

Relevance to current study. Though there is a limited literature linking

³ r/RoastMe is a subreddit where Redditors submit images of themselves for others to deride.

past user activity to longitudinal outcomes, the research in general suggested that information on prior activity may be helpful in predicting return for advice. For example, r/relationships users with a history of using Reddit as a source for help (as indicated by previous usage of other support and advice subreddits) may be more inclined to continue using r/relationships as another source of advice.

Metadata of community's response. Post score (n.b., a higher score means the post received more upvotes), number of comments the post received, and the timestamps of the post and comments have been used to represent support in prior studies. The number of comments and post score have been used as proxies for social support in a number of Reddit studies (e.g. De Choudhury & De, 2014; Pavalanathan & De Choudhury, 2015; Cunha, Weber & Pappa, 2017). This interpretation is akin to the concept of received support in the social support literature, which is simply the amount of support enacted (Tardy, 1985; Barrera, 1986).

This may be acceptable for post scores, since a higher post score means more users upvoted the post; as asserted by Van Mieghem (2011), the Reddit score is the collective evaluation of the community on the post. However, not everything articulated in a discussion about another's troubles is incontrovertibly supportive (Goldsmith, 2004); more comments are not necessarily better or indicative of quality support (Feng et al., 2016). Given the ambiguity of interpretation, the current study did not assume that a greater number of comments received corresponded to greater social support. The number of comments was taken only to be a measure of interest (whether positive or negative) by the community to the user's post.

As for the post and comment timestamps, they have been used to calculate the time lag between the time of posting to the time of receiving the first comment. Faster responses were observed on mental health subreddits versus non-mental health subreddits; this speed was interpreted as a demonstration of empathy towards the posting Redditors (Pavalanathan & De

Choudhury, 2015). The speed of response may play some role in affecting perceptions of support; female participants perceived emails received with a shorter delay period as providing higher quality support, though the effect size was small (Ledbetter, 2007). Nonetheless given that it is difficult to infer the degree of positivity of a comment just from its delivery speed, the current study did not assume that swifter comments were analogous to higher quality support. The time delay between the post and its first received comment was again interpreted only as an indicator of the community's interest.

Relevance to current study. Following the principles of conditioning, it has been hypothesised that positive responses received by the user following usage of an online tool would encourage them to persist in the activity (Davis, 2001). Receiving a reply has been found to be linked to a user returning to an online group (Arguello et al., 2006). Receiving more comments and a higher post score on one's post in a weight loss subreddit were linked to user return and the positive outcome of self-reported weight loss (Cunha et al., 2017). It was hence expected in the current study that a higher post score, more received comments, and a shorter time to receiving the first comment would be linked to a user returning to seek for advice. As described earlier, these metrics were conservatively interpreted as measures of community attention and not support.

Stylistic features of text. The stylistic features of text encapsulate *how* the message is expressed. The features examined in this study are length, readability, function word usage of posts and comments, and hedge, modal, and imperative use in comments.

General features: post length and readability. A basic feature of any message is its length. The length of writing has been construed as indicative of greater self-disclosure (Barak & Gluck-Ofri, 2007; De Choudhury, Kiciman, Dredze, Coppersmith & Kumar, 2016) as longer text may be due to the author describing their situation in detail. Greater trust in the other party is correlated with self-reported intent to disclose and intended amount to disclose (Wheless & Grotz, 1977). It was hence expected in the current study that longer initial

posts would be related to return as the users who wrote longer advice-seeking posts may also trust the forum more.

Longer writing may similarly be a demonstration of greater interest from the comment author as it can be interpreted as the author choosing to spend more time crafting their message to the recipient (Walther, 2007). Length has also been found to be related to perceived expertise in a study of online medical advice (Toma & D'Angelo, 2014) and to trustworthiness, when paired with context-appropriate language, in online consumer reviews (Filieri, 2016). It was thus expected that receiving longer comments would be related to returning for advice in future.

Another basic feature of a message is its readability – how easy it is to understand. Readability is commonly assessed in studies of online text (e.g. Gkotsis et al., 2016; Tan, Niculae, Danescu-Niculescu-Mizil & Lee, 2016; De Choudhury et al., 2016; Park & Conway, 2018). Lower readability was found in the writings of Redditors who eventually moved from mental health subreddits to post on r/SuicideWatch⁴ (De Choudhury et al., 2016). Redditors on mental health subreddits wrote posts that were less readable versus those on non-mental health subreddits (Park & Conway, 2018). These lower readability scores have been interpreted as an indication of a relatively weaker ability to express one's thoughts eloquently (De Choudhury et al., 2016; Park & Conway, 2018). At the outset of the current study, it was uncertain if the more active r/relationships advice-seekers would also overlap with the previously studied populations of Redditors using mental health subreddits. Thus, no specific relationship was defined between readability and returning for more advice on r/relationships.

Function words. Function words like pronouns, articles, and prepositions –grammatical words that string semantic content together – form the cornerstone of language (Gonzales, Hancock & Pennebaker, 2010). Though

⁴ r/SuicideWatch is a subreddit geared towards supporting Redditors thinking of committing suicide.

not as obviously meaningful as content words, function words are stable indicators of individual differences (Pennebaker, Mehl & Niederhoffer, 2003). Varying mental states and personality are related to the distinct ways that different people utilise function words (Boyd & Pennebaker, 2016), making them apt for studying user-level differences. There were no explicit expectations set for the relationships between function word categories and returning for more advice. The function word categories, as defined in the Linguistic Inquiry and Word Count (LIWC2015) software (Pennebaker, Boyd, Jordan & Blackburn, 2015) were included to check if and how the posts written by and comments received by returning advice-seekers and non-returning users differed.

One of the most prominent function word categories is pronouns. Pronouns are a useful gauge of where an individual's focus lies (Chung & Pennebaker, 2007). First-person singular pronouns (e.g. I) indicate self-focused attention (Chung & Pennebaker, 2007), whereas other pronouns like the first-person plural pronouns (e.g. we) emphasise a more social and group-based focus (Kane & Rink, 2015). First-person singular pronouns have been used as a cue for self-disclosure (i.e., the more first-person singular pronouns used, the more self-disclosure in text) (e.g. Harper & Harper, 2006; Walther, 2007; Barak & Gluck-Ofri, 2007; Brunet & Schmidt, 2007; De Choudhury & De, 2014; Rains, 2014). A high usage of first-person singular pronouns however might be symptomatic of high self-preoccupation and has consistently been found in the writings of individuals facing mental health distress (e.g. Stirman & Pennebaker, 2001; De Choudhury et al., 2016).

In the specific context of received advice comments, framing advice by disclosing some of the advisor's own personal experiences is a tactic used to signal the advisor's expertise and to provide advice in a less head-on manner (Morrow, 2006; Harrison & Barlow, 2009). On the other hand, third-person pronouns (e.g. they) co-occur often with past tense and the perfect aspect; together they suggest a more detached reporting-style narrative (Biber, 1986).

Given the rich amount of information provided by pronoun usage, the proportions of first person singular, first person plural, second person, third person singular, third person plural, and impersonal pronouns in posts and comments were calculated.

The other function word categories (i.e., articles, prepositions, auxiliary verbs, adverbs, conjunctions, and negations) may be helpful in getting more insight on the state of the individual at the time of writing too. Higher usage of articles and prepositions has been linked to categorical thinking (Pennebaker et al., 2014; Boyd & Pennebaker, 2016); prepositions are often used in academic writing as a way of incorporating more details (Biber, 1988). In contrast, greater usage of auxiliary verbs, conjunctions, and negations alongside pronouns are more characteristic of personal narratives (Pennebaker et al., 2014). Text produced within a limited timespan (e.g. an everyday conversation) tends to be less assiduous in structure due to the restricted time for message construction (Biber, 1988). Such text thus is typically fragmented, relying on conjunctions to loosely pull together units of thought (Biber, 1988). In summary, function words may be useful in identifying the differences in which the returning advice-seekers expressed themselves to the forum versus the non-returning advice-seekers, and the differences in which the forum responded to the two groups, if any.

Modals, hedges, and imperatives in comments. As mentioned earlier in this chapter (2.2.1 Definition of advice), advice may be threatening to the recipient's face (Goldsmith, 1999; Goldsmith & MacGeorge, 2000; Locher & Hoffman, 2006; Morrow, 2006; Tost et al., 2012; Feng et al., 2017). Advice that is constructed in a way that is careful of the recipient's face is assessed more favourably and is more helpful in assisting the recipient in coping (Feng & MacGeorge, 2010). In the construction of advice, modals (e.g. "You *might* want to do that.") are commonly used instead of bare imperatives (e.g. "Do that!") as imperatives may be viewed as too direct (Wierzbicka, 1985; Hudson, 1990; McDonald & Woodward-Kron, 2016). Imperatives can appear hostile to

the recipient's face as they may appear to not respect their autonomy (Brown & Levinson, 1987; Goldsmith, 1999; McDonald & Woodward-Kron, 2016).

Besides modals, advice may also be peppered with hedges (Brown & Levinson, 1987; Locher, 2006; Sillence, 2010) – words which highlight the subjectivity and openness of a statement (e.g. *possibly*, *perhaps*) and mitigate the certainty of an utterance (Biber, 1988; Hyland, 2005). Hedges are thus also helpful in maintaining the recipient's face (Brown & Levinson, 1987). In Collison et al.'s (2018) smaller study on r/relationships, it was found that commenters tended to use more discrepancy words (e.g. *should*, *could*) versus posters. Collison et al. (2018) remarked that this finding represented “greater cognitive processing” and “counterfactual thinking” by the commenters (p.307). However, in light of the studies on modal usage and hedging in advice, it may be argued that Collison et al.'s (2018) finding could also be interpreted as concerning face considerations.

For the current study, given the significance and frequency of use of mitigative linguistic strategies when giving advice in the Anglo-Western culture (Locher, 2006), it was expected that a greater usage of modals and hedges, and lesser use of imperatives in received comments would be linked to returning to seek for more advice.

Semantic features of text. Finally, the semantic features of text cover its content– *what* is being expressed in a message. The examined features in this study are topics, sentiment, and words related to time-focus and cognitive processes in posts and comments.

Topics discussed in posts and comments. Working directly with individual words to infer the subjects of discussion on the forum can be overly complicated; there are simply too many words to tackle in large datasets (Schwartz & Ungar, 2015). The current study thus focused on locating and comparing topics (i.e., groups of associated words) which are easier to interpret (Schwartz & Ungar, 2015). An unsupervised topic modelling algorithm (Latent

Dirichlet Allocation) was used to automatically create topics (see Chapter 3 and Appendix A).

Understanding the topics covered by the advice-seeking posts may be helpful in discerning the type of relational problem faced by the user. Previous research has found that the seriousness of the problem was an important variable related to the inclination to get help (Hinson & Swanson, 1993). A study of Internet usage in children and adolescents found that difficult life events (e.g., problems with parents, sickness, school issues) were related to using the Internet with the motivations of getting information and fulfilling social needs (Leung, 2007).

Furthermore, certain themes or topics covered in posts have been found to be associated with more response than others (Wang, Burke & Kraut, 2013; Davis et al., 2015). A qualitative study on a teen bulletin board which covered extremely similar themes to r/relationships (e.g. relationships, peers, sexuality) found similarly that different topics elicited different amounts of replies (Suzuki & Calzo, 2004). In the current study, it was hence expected that some post topics might be observed to be related to returning for more advice. However, given the unsupervised method of creating the topics, no specific hypotheses were made given that the topics would not have been known prior to running the topic modelling algorithm.

Topics were also created for the comments on each post. Linguistic features in received comments like mentions of family have been found in prior research to be related to decreases in user distress on an online mental health support forum (Soldaini, Walsh, Cohan, Han & Goharian, 2018). Similar to the post topics, it was expected that some topics covered in the comments may be related to returning to seek for advice. Again, given that the exact topics were not known prior to the study, no specific hypotheses were made.

Sentiment and emotion in text. Getting a sense of the sentiment and emotion in the text – especially of the negative variety – can provide further information on the state of the individual at the point of writing. Words related

to negative sentiment and emotion have been found in previous studies to be more heavily utilised by depressed Twitter users in their tweets (Park, Cha & Cha, 2012), depressed Redditors (Wolohan, Hiraga, Mukherjee, Sayyed & Millard, 2018), and by distressed students in their Facebook posts (Barazova, Choi, Whitlock, Cosley & Sosik, 2017). Importantly, anxiety has also consistently been linked to the decision to seek advice (Gino et al., 2012; MacGeorge et al., 2016a).

Emotions and the sentiment in text also have implications for the response of others to the poster. Negative emotion has consistently been found to be an important feature in eliciting response online – it was a significant feature in Althoff et al. (2014) in predicting successful requests on the subreddit Random Acts of Pizza (i.e., posts which successfully received their requested pizza) and was related to higher levels of activity on BBC forums (Chmiel et al., 2011). Conveying negative emotion in appropriate circumstances has been linked to the willingness of others to help and greater evocation of support (Graham, Huang, Clark & Helgeson, 2008). Highly emotional posts on an online social support board for language assistants tended to generate responses that were empathetic in nature (Ruble, 2011). Similarly, in a laboratory study of online support, participants rated posts that shared more of the writer's emotions as more credible (Li et al., 2019). Participants were also more likely to report intention to provide emotional support to the writer (Li et al., 2019). In the current study, negative and positive sentiment were measured, and proportion of words relating to anxiety, anger, and sadness calculated. It was expected that higher levels of negative sentiment and greater usage of anxiety-related words would be related to returning for advice.

Similar calculations were done for the comments received for each post. In a smaller scale study of r/relationships, Collison et al. (2018) found that commenters tended to display more positive emotions than posters, but no differences were found for negative emotions. This finding was regarded as representative of the socially supportive position the commenters undertook

(Collison et al., 2018). It was thus expected that higher levels of positive sentiment in the comments received may be related to returning for advice.

Time focus and cognitive processes in posts and comments. Words related to an individual's time focus and cognitive processes are two categories defined by LIWC2015 (Pennebaker et al., 2015) which may be useful in further characterising repeat advice-seekers. Words related to cognitive processes (e.g. 'know', 'cause') – though not directly related to the topic of the post – are more semantically meaningful than function words (Toma & D'Angelo, 2015) and are related to cognitive reasoning (Pennebaker et al., 2015). Time focus – words indicating the writer's relative focus on the past, present, or future (Pennebaker et al., 2015) – may also be applicable to the study. Collison et al. (2018) found that typically advice-seekers on r/relationships showed more attention to the past, whereas advice-givers paid more attention to the present and future. Depressed Redditors have been found to focus more on the present and future (Wolohan et al., 2018).

The proportion of words used in the posts and comments received that appear in the LIWC2015 subcategories for cognitive processes (i.e., insight, causation, discrepancy, tentative, certainty, differentiation) and time focus (i.e., past, present, future) were calculated. However, no specific relationships were expected for these word categories. As with the function words, they were included to assess if they could provide further information to distinguish between advice-seeking returnees and non-returnees.

CHAPTER III: OVERVIEW OF R/RELATIONSHIPS AND ITS USERS

3.1 Overview of chapter

This chapter provides a quantitative overview of the content of r/relationships and the behaviour of its users. It also covers in brief the process of generating this overview. All work was done in Python 3.

3.2 Purpose of overview

As explained in Chapter 2, the focal goal of the current study was to explore how (1) user-level differences and (2) differences in the forum community reception may be associated with a user deciding to seek advice again on r/relationships. Before proceeding with the main research question, the project attempted to get a sense of the topics discussed on the forum and the past activity of its users. This overview was key for two reasons. Firstly, unlike the well-studied mental health subreddits, there have been almost no other academic studies conducted on this forum, save for Collison et al.'s (2018) study on romantic-only submissions. However, r/relationships has a much wider scope than just romantic relationships. It was thus prudent to start with a better understanding of the scope of issues discussed on the forum of study and its users. Secondly, the variables created from this overview were to serve as quantitative representations of the topics covered and activity of the r/relationships Redditors for use in the analyses in Chapter 4.

The next section (3.3 Data collection) covers in brief the process of data collection. After which, the topic modelling process to retrieve the topics discussed by the posters and commenters on r/relationships is described and its results presented. This chapter closes with the process and results of summarising the past Reddit activity of the posters via a community detection algorithm and text clustering.

Figure 2. *Flow of analyses covered in Chapter 3.*

3.3 Data collection

Reddit has an application programming interface (API) that is freely available for use by the public. A user can send a request to the API to retrieve various post, comment, and user details.

3.3.1 Collection of posts and comments. In March 2018, Reddit removed timestamp-based search functionality as it shifted to a new search system (“Update To Search API : changelog”, 2018). As a workaround, Pushshift, which independently maintains a database of all content posted on Reddit and allows for timestamp-based search, was used to retrieve the needed post IDs. A list of IDs of posts made from 1 March 2017 00:00:00 GMT to 31 March 2018 23:59:59 GMT was retrieved using the Pushshift API. A preliminary removal of post IDs of posts that had been removed by the r/relationships moderators was done. The final list of post IDs was then used to retrieve the post information and their related top-level (direct) comments from the Reddit API. The API pull resulted in 145144 posts and 928051 direct comments.

From the 145144 posts, one post had a formatting issue which resulted in its text being prematurely cut off. It was dropped. Posts that were deleted by the user (hence having only [deleted] as body text) and posts with below 80 characters were removed. This figure was decided on via manual inspection – an example of such a post is “... delete this please :)”. 70520 posts (hereon referred to as the 70520-dataset) and 513591 corresponding comments remained for analysis.

3.3.2 Collection of user history. User history could only be retrieved for posters with available usernames. The usernames of the post authors from the 70520-dataset were filtered accordingly:

- (1) usernames that were simply [deleted] or None were removed (no available information/deleted accounts), and
- (2) moderators of r/relationships were removed.

For the authors with available usernames ($n=60434$), the timestamps of their first posts in the 70520-dataset was retrieved. This information was used in the API request to Pushshift. All available posting and commenting activity before and after their first posts in the dataset were collected.

3.4 Topics discussed on the forum

3.4.1 Methods. The topic modelling algorithm used was Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003). The main input required for LDA is the number of topics, k , that the algorithm should produce (Moessner et al., 2018). Derived topics are made up of consistently co-occurring words (Schwartz & Ungar, 2015; Carron-Arthur, Reynolds, Bennett, Bennett & Griffiths, 2018). LDA can be applied to all the forum posts to automatically derive the k main topics of discussion on the forum (Blei, 2012). This is opposed to relying on hard-to-scale qualitative methods and/or predetermined word lists which may miss less conventional topics in the data (Stanton, Boyd, Pulverman & Meston, 2015; Kern et al., 2016). Another benefit of LDA is its ability to calculate the proportions of each topic in every forum

post (Carron-Arthur et al., 2018). Such representations of the posts allow for post content to be quantitatively compared.

As the comments may bring up topics not covered in the original posts (e.g. seeking help from the authorities), the posts and comments were modelled separately. The 513591 comments were grouped by the posts they were replying to, resulting in 68925 comment groups. Topic modelling was performed with Gensim (Rehurek & Sojka, 2010). 28 topics were retrieved for the 70520 posts. 36 topics were retrieved for the 68925 comment groups. Details of the modelling can be found in Appendix A.

3.4.2 Topics identified. Having selected the final number of topics, the next step was to manually label the topics. LDA by default returns the most representative (i.e., highest probability) words for each topic which can be used to guide topic labelling (Chancellor, Lin, Goodman, Zerwas & De Choudhury, 2016; Moessner et al., 2018). To further assist in topic interpretation, the relevance metric (Sievert & Shirley, 2014) was used to reorder the words, such that high probability words in a topic which were *not* also high probability words in other topics were ranked higher. The keywords used in labelling the topics for posts and comment groups are in Appendix A.

The 28 topics derived from posts are: (1) general forum language, (2) hedging/quantifiers, (3) past partners, (4) communication, (5) romantic feelings, (6) fighting/arguments, (7) mental health, (8) living arrangements: hygiene/chores, (9) texting/calling, (10) moving, (11) dating, (12) social media/trust issues, (13) family, (14) finances, (15) socialising, (16) friends, (17) school, (18) marital, (19) living arrangements: space, (20) sexual, (21) long distance relationships, (22) jobs, (23) trips/occasions, (24) drinking/drugs/religion, (25) hobbies/interests, (26) body image, (27) miscellaneous, (28) getting married.

The 36 topics derived from comment groups are: (1) general forum language, (2) communication, (3) cutting contact/poor behaviour, (4) romantic: long-term/long distance, (5) interpretation of issues, (6) taking action, (7)

mental health: therapy, (8) family, (9) marriage, (10) activity scheduling, (11) physical attractiveness, (12) dating, (13) finances, (14) food/eating preparation, (15) careers/school, (16) mental health: disorder/diagnosis/professional, (17) friends: current issues, (18) hobbies/interests, (19) age gap, (20) lying/secrets, (21) sleeping arrangements, (22) sex, (23) friends: making, (24) family: moving/extended/loss, (25) cheating, (26) special occasions, (27) electronic communication, (28) social media/app usage, (29) abuse, (30) love compatibility, (31) space/touch, (32) lease, (33) pregnancy/birth control, (34) gaming, (35) drinking/drugs, (36) miscellaneous.

3.4.3 Summary of results. It was observed that the subreddit spanned a diverse range of topics. Though romantic feelings were a mainstay topic, users also came to the subreddit looking for advice on more delicate issues like marriage, mental health, and finances. This confirmed that the spread of topics that people seek advice for online is far broader than previously tested in researcher-created scenarios rely on undergraduate samples (e.g. Feng et al., 2016; Rains et al., 2018). Cohen and Wills (1985) previously noted the need for more diverse samples with greater variabilities of stress and support. The diverse topics found by LDA also support the utility of the current study in furthering previous work on advice by way of the naturalistic data allowing for an anonymised exploration of topics that may otherwise be dropped from questionnaires due to sensitivity (e.g. abuse in Leung, 2007).

Similar to other research that modelled comments separately from posts (Pappa et al., 2017), it was found that the topics discussed in the comments were relatively similar to those covered by the posts. This is intuitive since the comments were direct replies to the posts. Still, modelling the comments separately was preferable – commenters brought up topics (e.g. cutting contact) that were far less/not mentioned in the original posts.

As mentioned earlier (see 3.2 Purpose of overview), the topic proportions provided by LDA for the posts and comments were used later as variables in statistical analysis and input for predictive modelling (see Chapter

4). Other studies have used LDA in a similar manner (e.g. Pappa et al., 2017; Ji, Ping, Fung, Pan and Long, 2018). However, not all topics were helpful for this endeavour. Again, the purpose of conducting LDA in this study was to retrieve a quantitative representation of post and comment *content* – but LDA is incapable of differentiating between writing style and content (Wang et al., 2013) (e.g. general forum language topic versus content topics like fighting/arguments). LDA also does not guarantee coherent topics and may give rise to junk topics (Carron-Arthur et al., 2016) – it does not account for structure or syntax in modelling (Guo, Barnes & Jia, 2017). LDA topics that were not clearly about post content were thus dropped from further analysis. The dropped topics for posts were: (1) general forum language, (2) hedging/quantifiers, and (3) miscellaneous. The dropped topics for comments were: (1) general forum language, and (2) miscellaneous.

3.5 Summarising posting history of r/relationships users

While quantifying post and comment content was a straightforward task, summarising the posting activity of every identifiable user ($n=60434$) in the year preceding their first post in the 70520-dataset was far more challenging. Ideally, each subreddit would act as a count variable (e.g. four posts and seven comments in r/suicidewatch would give the user a count of 11 on a r/suicidewatch variable) (e.g. in Tamersoy et al., 2015). Unfortunately, this was impractical with the huge number of subreddits. Previous studies that incorporated posting history into their analyses tackled this by manually selecting a handful of popular subreddits related to the domain of study (e.g. mental health) (Tamersoy et al., 2015; Kasunic & Kaufman, 2018; Li, Mihalcea & Wilson, 2018). However, posters on r/relationships seek advice on a very broad range of topics. It thus was difficult to pinpoint exactly which subreddits to include in the analysis. Hence, instead of manually identifying specific subreddits to include, the current study attempted to identify thematically-relevant *groups* of subreddits. In other words, subreddits were clustered according to their discussion themes and manual selection was performed on

these clusters instead. For example, wedding-related subreddits were clustered together and acted as one count variable, as opposed to taking r/wedding and r/weddingplanning as two separate count variables. Note that r/relationships itself was not included in this clustering; posts and comments on r/relationships from the preceding year were separately counted.

3.5.1 Methods. Subreddits were initially clustered together by theme by applying the Louvain community detection algorithm (Blondel, Guillaume, Lambiotte & Lefebvre, 2008) to a network graph of subreddits. In the graph, subreddits were nodes and linked if they shared at least one r/relationships poster from the current study's dataset; the links were weighted by the number of r/relationships posters the certain pair of subreddits shared. The Louvain community detection algorithm does not require manual specification of the number of clusters to find; it works by identifying dense subgroups within the graph (Schaub, Delvenne, Rosvall & Lambiotte, 2017). The assumption made here was that well-connected subreddits that shared many common users should be thematically similar. This process of network creation and community detection analysis was heavily inspired by Olson and Neal's (2015) work in applying the Louvain algorithm to the posting histories of over 800,000 Redditors to create a graph of topical interests.

The Louvain algorithm returned 31 groups of subreddits (full listing in 3.5.2 Groups of subreddits identified). A number of groups stood out as potential groups of interest (also see 3.5.2 Groups of subreddits identified). From these potential groups, there was a particularly large group of 185 subreddits that had the common theme of being more personal – for example, r/dating_advice, r/bisexual, r/SuicideWatch. As the 185 subreddits were still diverse topically despite sharing a common theme – spanning from jobs to dieting to mental health – the group was further refined into smaller groups based on topical similarity. This was done by clustering these subreddits based on their text (e.g. Sharma & De Choudhury, 2018). By clustering subreddits based on their word usage, topically similar subreddits were grouped together.

The text clustering algorithm used was non-negative matrix factorization (NMF) as implemented in scikit-learn (Pedregosa et al., 2011). Similar to LDA, NMF requires specification of k (the number of clusters to retrieve). From this single large group of 185 subreddits, 44 groups of subreddits were created; a number of potential subreddit groups of interests were again manually picked out for further study. Full details of the network analysis and text clustering are in Appendix B.

3.5.2 Groups of subreddits identified. The 31 groups obtained from running the Louvain algorithm were manually labelled based on the subreddits that comprised each group. The groups were relatively diverse – spanning from subreddits covering memes/drama, to gaming, sports, and cryptocurrency. Among the 31, six groups consisted of subreddits containing content that may be relevant indicators of a user choosing to seek advice online often (e.g. suggestive of personal troubles, relational issues, and/or prior help-seeking). These six groups were selected for further analysis; they were: (1) drug/crime-related subreddits, (2) transgender persons-related subreddits, (3) parenting-related subreddits, (4) gay men-related subreddits, (5) self-harm-related subreddits, and (6) personal/sharing-related subreddits. More details on all 31 subreddit groups can be found in Appendix B.

As mentioned in the previous section (3.5.1 Methods), the personal/sharing-related subreddits group was extremely large and was refined into 44 smaller groups using text clustering. The topical diversity was clear when examining the 44 derived groups – they covered a broad range of topics, from weddings, pets, mental health, to dieting. From the 44 groups, 18 groups were selected for further analysis on the similar basis of them possibly being relevant for identifying returning advice-seekers. The 18 selected groups were: (1) weight-related subreddits, (2) validation-related subreddits, (3) dating-related subreddits, (4) wedding-related subreddits, (5) relationship-related subreddits, (6) sex-related subreddits, (7) advice/help-related subreddits, (8) sexual orientation-related subreddits, (9) JustNo-themed subreddits, (10)

confession-related subreddits, (11) mental health-related subreddits, (12) anxiety-related subreddits, (13) incel culture-related subreddits, (14) drinking-related subreddits, (15) subreddits focusing on parental figures, (16) the JustEngaged (engagement) subreddit, (17) the NoFap subreddit, and the (18) Meditation subreddit.

In total, 23 groups of subreddits were selected for further analysis. For every r/relationships poster, their posts and comments in the year preceding their first post in the 70520-dataset to subreddits in each group were summed. In other words, these 23 count variables were used to represent each user's one-year posting history to subreddits of interest in a succinct manner.

3.5.3 Summary of results. By using the community detection process outlined by Olson and Neal (2015) to identify thematically-similar interest groups in conjunction with text clustering, the current study was able to get a broad overview of each identifiable r/relationships poster's past activity. Though more crude than manually selecting specific subreddits as done in previous research, this process circumvented the issue of having to manually sift through the 24063 subreddits present across the histories of all r/relationships posters⁵.

It is clear that together, the r/relationships users had a diverse posting history. 23 subreddit groups stood out as possibly suggestive of users who view Reddit as a source of help or support for their personal issues. As mentioned in Chapter 2, past online activity may inform future behaviour, and so these 23 subreddit groups were marked for further analysis to see if they were related in any way to returning for advice (see Chapter 4).

⁵ A quick check showed that following Tamersoy et al. (2015)'s methodology in this case would result in the omission of several subreddits of interest – like the self-harm subreddits and r/mentalhealth.

CHAPTER IV: FACTORS RELATED TO REPEAT ADVICE-SEEKING

4.1 Overview of chapter

This chapter covers the methods and results of the comparisons between repeat and non-repeat advice-seekers on r/relationships. All analyses were focused on identifying possible (1) user-level differences and (2) differences in forum community reception that may be associated with a user deciding to seek advice again on r/relationships. All work was done in Python 3 except where indicated.

The chapter starts by covering how users were selected for the analyses (4.2 User selection). Following that, a list of all variables examined – inclusive of those generated from Chapter 3 – is provided in 4.3 Variables of study. Sections 4.4 to 4.7 cover the main analyses of this study. Section 4.4 covers the base comparison between repeat and non-repeat advice-seekers. Section 4.5 describes the same comparisons between repeat and non-repeat advice-seekers, but split and performed separately for male and female users. Section 4.6 introduces a multiclass problem where the additional behaviour of advice-giving is accounted for alongside advice-seeking. The chapter closes with section 4.7, which compares a different sample of repeat advice-seekers (i.e., ‘veteran’ repeat advice-seekers) versus non-repeat advice-seekers.

4.2 User selection

4.2.1 Removing throwaways. Not all 60434 identifiable r/relationships posters were used for the analysis. Throwaways (i.e., one-time-use accounts which were created to post once on r/relationships and never used again) were removed from the sample. Previous studies have shown that these accounts are linguistically different and received differently from non-throwaways (De Choudhury & De, 2014; Pavalanathan & De Choudhury, 2015).

4.2.2 Identifying repeat r/relationships advice-seekers. Repeat advice-seekers were posters who created a new advice-seeking post within three months after their first post in the 70520-dataset. Thus, to identify repeat

advice-seekers, the activity of each poster in the three months following their post in the 70520-dataset was checked for r/relationships posts. Following this, 6162 posters were picked out.

Of these 6162 posters, it was possible that some returned simply to offer thanks to the forum and/or update the forum on the conclusion of their initial problem⁶. To ensure a high confidence that the returning user group contained purely repeat advice-seekers, the subsequent r/relationships posts of these 6162 users were manually checked. The user was tagged as a repeat advice-seeker if:

- (1) they had at least one r/relationship post within the next three months without an update flair (i.e., they were bringing up a new problem to the community), or
- (2) the user had only post(s) with an update flair within the next three months, but the post(s) had available post text showing that the user was still explicitly seeking for more advice on the problem.

Following the above criteria, 5502 repeat advice-seekers were established⁷.

Repeat advice-seekers with no prior r/relationships posts. Some of the 5502 users had posted on r/relationships prior to the 70520-dataset. For the first three sets of analyses (i.e., base analysis, accounting for gender, accounting for advice-seeking), only returning users who had never posted on r/relationships prior to the 70520-dataset were selected. This sampling decision was similar to that used in other studies to control for the effects of previous experiences when examining longitudinal user behaviour (e.g. Chang & Danescu-Niculescu-Mizil, 2019). This decision was bolstered by past studies which have found that prior postings on a community already strongly predict future postings on the same community (e.g. Arguello et al., 2006). Hence, users were selected as

⁶ r/relationships allows one update for each problem posted. It also stipulates that such thank you/update posts be flagged with 'Update' in the title and tagged with the 'Updates' flair.

⁷ Automating the process by relying on post titles/flairs would not have drastically changed the selected sample. Out of the 660 dropped users, 96.5% had 'Update' in their post titles and 87.7% had 'Updates' as a post flair. Out of the 5502 selected users, only 4% had 'Update' in their post titles and 4% had 'Updates' as a post flair.

repeat advice-seekers for the first three sets of analyses if (illustrated in Figure 3 as well):

- (1) they had never sought for advice before on r/relationships (i.e., never created a post on r/relationships prior to their first post in the 70520-dataset), and
- (2) in the case of the user actually having no activity on Reddit prior to their first post in the 70520-dataset, the user had to display activity beyond commenting and/or posting on r/relationships after their first post. This condition was applied to avoid including accounts which had the sole purpose of acting as an r/relationships throwaway.

After filtering, 2539 repeat advice-seekers were retained. Users who did not receive comments on their first r/relationships post were dropped, leaving 2469 repeat advice-seekers. Lastly, users with no information on age and/or gender⁸ were dropped, leaving 2371 repeat advice-seekers for use in the first three analyses.

⁸ Gender was taken as binary (male/female). Non-binary/genderfluid users were dropped due to the small sample size (n=20). Transgender individuals were coded as the gender they identified as.

Figure 3. *Selecting repeat advice-seekers with no prior r/relationships posts.*

Repeat advice-seekers with prior r/relationships posts. For the last set of analyses (i.e., ‘veteran’ repeat advice-seekers versus non-repeat advice-seekers), repeat advice-seekers who posted on r/relationships prior to the 70520-dataset were analysed. To retrieve them from the 5502 repeat advice-seekers, the following conditions were applied (illustrated in Figure 4 as well):

- (1) the user had sought for advice (i.e., created a post) before on r/relationships in the three months before their first post in the 70520-dataset⁹, and

⁹ r/relationships automatically deletes posts that do not fit its formatting criteria. Some users have histories filled with failed attempts (and re-attempts) at formatting their posts correctly. To avoid mistakenly labeling these as advice-seeking instances, a post was only considered as a separate advice-seeking instance if there was at least a two-day gap between it and the next r/relationships post by the user.

(2) the user had activity beyond commenting and/or posting on r/relationships. Those who failed this check were dropped. Again, this was done to avoid including r/relationships throwaway accounts.

After filtering, 852 repeat advice-seekers were retained. Users with no comment data were dropped, leaving 831 repeat advice-seekers. Lastly, users with no information on age and/or gender were dropped, leaving 788 repeat advice-seekers for use in the last set of analyses.

Figure 4. *Selecting repeat advice-seekers with prior r/relationships posts.*

4.2.3 Identifying non-repeat r/relationships advice-seekers. To ensure that the non-repeat group consisted of users with no involvement in r/relationships prior to and after their first post on r/relationships, the following conditions were applied (illustrated in Figure 5 as well):

- (1) the user had never sought for advice on r/relationships before (i.e., never posted on r/relationships prior to their post in the 70520-dataset), and
- (2) the user had to do more than comment only on r/relationships after their post in the 70520-dataset. Those who failed this check were dropped.

After filtering, 16588 non-repeat advice-seekers were retained. Users with no comment data were dropped, leaving 16164 non-repeat advice-seekers. Lastly, users with no information on age and/or gender were dropped, leaving 15525 non-repeat advice-seekers. These 15525 non-repeat advice-seekers were used for the first three sets of analyses.

Figure 5. *Selecting non-repeat advice-seekers.*

For the last set of analyses (with the ‘veteran’ repeat advice-seekers), given that all ‘veteran’ repeat advice-seekers definitely had activity on Reddit prior to their first post in the 70520-dataset, only non-repeat advice-seekers with Reddit activity prior to their post in the 70520-dataset were retained. This left 11429 non-repeat advice-seekers for the last set of analyses.

4.3 Variables of study

Table 1 summarises all 149 variables (grouped according to the type of information they provide) created from the data of the r/relationships posts, comments, and its users. The creation of the subreddit clusters (under activity information) and LDA topics (under semantic content of posts/comments) was

covered in Chapter 3 and Appendices A and B. Information on the creation of the other variables can be found in Appendix C.

Table 1. *149 variables used in current study.*

Variable type	Variables
User information	Self-reported age of user at first r/relationships post in 70520-dataset
	Self-reported identified gender of user at first r/relationships post in 70520-dataset
Account information	Account age at time of first r/relationships post in 70520-dataset
	Was r/relationships post in 70520-dataset the user's first activity ever on Reddit?
Activity information	Total count of all posts and comments made on Reddit in year prior to first post in 70520-dataset
	Total count of all posts and comments made on 23 subreddit clusters of interest in year prior to first post in 70520-dataset (from Chapter 3): (1) drug-related, (2) transgender persons-related, (3) parenting-related, (4) gay men-related, (5) self-harm-related, (6) weight-related, (7) validation-related, (8) dating-related, (9) wedding-related, (10) relationship-related, (11) sex-related, (12) advice/help-related, (13) sexual orientation-related, (14) JustNo-themed, (15) confession-related, (16) mental health-related, (17) anxiety-related, (18) incel culture-related, (19) drinking-related, (20) subreddits focusing on parental figures, (21) the JustEngaged (engagement) subreddit, (22) the NoFap subreddit, (23) the Meditation subreddit
	Number of r/relationships comments in year prior to first r/relationships post in 70520-dataset
Stylistic information of user's first post in 70520-dataset	Readability of post
	Function word usage: (1) first-person singular pronouns, (2) first-person plural pronouns, (3) second-person pronouns, (4) third-person singular pronouns, (5) third-person plural pronouns, (6) impersonal pronouns, (7) articles, (8) prepositions, (9) auxiliary verbs, (10) adverbs, (11) conjunctions, (12) negations
	Length of post

Semantic information of user's first post in 70520-dataset	LDA topics (from Chapter 3): (1) Body image, (2) Communication, (3) Dating, (4) Drinking/drugs/religion, (5) Family, (6) Fighting/arguments, (7) Finances, (8) Friends, (9) Getting married, (10) Hobbies/interests, (11) Jobs, (12) LDR, (13) Living arrangements: hygiene/chores, (14) Living arrangements: space, (15) Marital, (16) Mental health, (17) Moving, (18) Past partners, (19) Romantic feelings, (20) School, (21) Sexual, (22) Social media/trust issues, (23) Socializing, (24) Texting/calling, (25) Trips/Occasions
	Time focus: (1) focus on past, (2) focus on present, (3) focus on future
	Cognitive processes: (1) insight, (2) causation, (3) discrepancy, (4) tentative, (5) certain, (6) differences
	Sentiment/emotion of post: (1) negative sentiment, (2) positive sentiment, (3) usage of anxiety words, (4) usage of anger words, (5) usage of sadness words
Metadata information of user's first post in 70520-dataset	Post score
	Total number of comments
	Seconds elapsed between post and its first comment
Stylistic information of top-level comments received on user's first post in 70520-dataset	Readability of top-level comments received
	Proportion of imperative sentences in received top-level comments
	Proportion of words that are modals/hedges in received top-level comments
	Function word usage in received top-level comments: (1) first-person singular pronouns, (2) first-person plural pronouns, (3) second-person pronouns, (4) third-person singular pronouns, (5) third-person plural pronouns, (6) impersonal pronouns, (7) articles, (8) prepositions, (9) auxiliary verbs, (10) adverbs, (11) conjunctions, (12) negations
	Median length of received top-level comments

Semantic information of top-level comments received on user's first post in 70520-dataset	LDA topics (from Chapter 3): (1) Abuse, (2) Activity scheduling, (3) Age gap, (4) Careers/school, (5) Cheating, (6) Communication, (7) Cutting contact/poor behavior, (8) Dating, (9) Drinking/drugs, (10) Electronic communication, (11) Family, (12) Family: moving/extended/loss, (13) Finances, (14) Food/Eating Preparation, (15) Friends: issues, (16) Friends: making, (17) Gaming, (18) Hobbies/interests, (19) Interpretation of issues, (20) Lease, (21) Love compatibility, (22) Lying/secrets, (23) Marriage, (24) Mental Health: disorder/diagnosis/professional, (25) Mental Health: therapy, (26) Physical attractiveness, (27) Pregnancy/birth control, (28) Romantic: long-term/long-distance, (29) Sex, (30) Sleeping arrangements, (31) Social media/app usage, (32) Space/touch, (33) Special occasions, (34) Taking action
	Time focus: (1) focus on past, (2) focus on present, (3) focus on future
	Cognitive processes: (1) insight, (2) causation, (3) discrepancy, (4) tentative, (5) certain, (6) differences
	Sentiment/emotion of received top-level comments: (1) negative sentiment, (2) positive sentiment, (3) usage of anxiety words, (4) usage of anger words, (5) usage of sadness words

4.4 First analysis: repeat vs non-repeat advice-seekers

The first set of analyses addressed the base question of: what distinguished the users who returned for advice from the users who did not? For this analysis, 17896 users were used: 2371 repeat advice-seekers and 15525 non-repeat advice-seekers (13.2% repeat advice-seekers). 80% of the data was used for the training set, and 20% for the testing set. Group proportions were kept roughly equal in both sets. The statistical tests and model building were done only on the training set. The testing set was held-out to evaluate the performance of the trained model.

4.4.1 Statistical tests. Based on a chi-square test of independence, there was no association between the user's first post on Reddit being their first post in the 70520-dataset and whether they returned, $X^2(1) = .385, p = .535$. However, there was a significant association between gender and whether the user returned, $X^2(1) = 33, \Phi = .048, p < .001$.

Mann-Whitney U-tests were performed with the Holm-Bonferroni correction¹⁰ ($\alpha = .05$) for multiple comparisons (147 comparisons) for the remaining continuous variables. Non-parametric tests were performed due to the skewness of multiple variables (e.g. post score). Only significant results kept after the Holm-Bonferroni correction are reported in Table 2.

Table 2. *Significant results after Holm-Bonferroni correction (repeat vs non-repeat advice-seekers).*

Variable	Group higher in mean ranks	Cohen's d
Self-reported age of user at first post in 70520-dataset	Non-repeat advice-seekers	.095***
Account age at time of first r/relationships post in 70520-dataset	Non-repeat advice-seekers	.164***
LDA comment topic: Communication	Repeat advice-seekers	.074***
LDA comment topic: Mental health: therapy	Non-repeat advice-seekers	.063***
LDA post topic: Communication	Repeat advice-seekers	.060***
LDA post topic: Dating	Repeat advice-seekers	.076***
LDA post topic: Mental health	Non-repeat advice-seekers	.072***
LDA post topic: Texting/calling	Repeat advice-seekers	.097***
Length of post	Non-repeat advice-seekers	.069***
Usage of conjunctions	Repeat advice-seekers	.083***
Comments on r/relationships for one year prior to first post in 70520-dataset	Repeat advice-seekers	.071***

¹⁰ The formula for Holm-Bonferroni is $\frac{\alpha \text{ for testing}}{\text{number of tests} - \text{rank number of pvalue} + 1}$. In this scenario, $\alpha = .05$ and number of tests = 147. After performing all 147 tests, the p -values for each test were ranked in ascending order (smallest to largest). The first (i.e., smallest p -value) was compared to $\frac{.05}{147 - 1 + 1}$, or $\alpha/147 = .0003401$. Since the p -value was smaller than .0003401, the null hypothesis was rejected. For the second smallest p -value, it was compared to $\frac{.05}{147 - 2 + 1}$, or $\alpha/146 = .0003425$. This process was repeated until the first p -value larger than the corrected α was reached (i.e., the first null hypothesis that cannot be rejected).

Activity on dating-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.036***
Activity on relationship-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.055***
Activity on advice/help-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.041***
Activity on anxiety-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.027***

*** $p < .001$

4.4.2 Prediction task. As mentioned in Chapter 2 (2.4.2 General considerations), to ensure the robustness of the findings, predictive models were built to support the findings of the statistical tests. The prediction task for the first set of analyses was a binary one with the repeat advice-seeker group set as the positive class (1) and the non-repeat advice-seeker group as the negative class (0).

Model training and selection. Using the scikit-learn library, L1-penalised logistic regression models were created (Pedregosa et al., 2011). Regularisation with L1 was used to avoid overfitting due to the large number of variables. The L1 penalty achieves this by creating a sparse model by limiting the magnitude of the regression coefficients; coefficients of less useful prediction variable(s) are pushed to zero, effectively eliminating the variable(s).

The sets of variables as specified in Table 1 (e.g. user information, activity information) were first evaluated individually. Following which, the variable sets were added incrementally to the model based on their individual predictive utility. Performance was evaluated at each step. Note that performance for all models was calculated by averaging over the results of 5-cross-validation folds (i.e., shuffle the data, use 4/5 of the data for training, 1/5 for evaluating; this is done five times) of the training set. The best performing model was selected. It was then trained on the full training dataset and its

performance on the testing set evaluated. Refer to Appendix D for full details on model training.

Results from best performing model. The best performing model was a combination of user information, account information, activity information, post LDA topics and post stylistic information (full performance metrics of model on testing set in Appendix D). For clarity, the model's performance on the testing set is displayed in a normalised confusion matrix (Table 3).

Table 3. *Normalised confusion matrix for model from first set of analyses.*

	Predicted as non-repeat advice-seeker	Predicted as repeat advice-seeker
Actual non-repeat advice-seeker	0.58	0.42
Actual repeat advice-seeker	0.41	0.59

59% of the repeat advice-seekers were correctly identified as such, and 58% of non-repeat advice-seekers correctly identified. Comparing against a dummy classifier which outputs predictions based on the class proportions of the training set (Pedregosa et al., 2011) (i.e., it would guess that most cases were the majority non-repeat class), the dummy classifier guessed 86% of the non-repeat advice-seekers correctly, but only got 16% of the repeat advice-seekers correct.

To get a sense of the variables related to predicting if a user is a repeat/non-repeat advice-seeker, the non-zero coefficients of the final model were examined. The full table of coefficients can be found in Appendix D. Table 4 shows the variables positively and negatively related to repeat advice-seekers.

Table 4. *Variables positively/negatively related to repeat advice-seekers.*

Positively related to repeat advice-seekers	Negatively related to repeat advice-seekers
Total count of account activity in preceding year	Age of user
Prior comments on r/relationships in preceding year	Account age

Activity on dating-related subreddits in preceding year	Activity on wedding-related subreddits in preceding year
Activity on relationship-related subreddits in preceding year	Activity on sexual orientation-related subreddits in preceding year
Activity on advice/help-related subreddits in preceding year	Body image (post LDA)
Activity on confession-related subreddits in preceding year	Family (post LDA)
Activity on anxiety-related subreddits in preceding year	Jobs (post LDA)
Activity on meditation-related subreddits in preceding year	Marital (post LDA)
Communication (post LDA)	Mental health (post LDA)
Dating (post LDA)	School (post LDA)
Past partners (post LDA)	Trips/occasions (post LDA)
Romantic partners (post LDA)	Length of post
Sexual (post LDA)	Usage of prepositions
Texting (post LDA)	
Usage of first-person singular pronouns (e.g. 'I')	
Usage of third-person singular pronouns (e.g. 'she')	
Usage of conjunctions	
Gender of user	

Summary of results. The results of the prediction model follow those of the statistical tests closely, providing support for their robustness. Repeat advice-seekers on r/relationships tended to be female and younger, with younger and more active Reddit accounts at the time of posting. Activity-wise, repeat advice-seekers on r/relationships tended to have previously made greater use of Reddit as a source of help (e.g., advice/help-related subreddits, anxiety-related subreddits, other relationship-related subreddits). They were also participating in discussions of the problems of others on r/relationships prior to

their first advice-seeking attempt on the forum. Repeat advice-seekers also were more focused on dating and the typical issues that surround it (e.g. communication, past partners, sex, and texting) instead of married life or family- and job-centric issues. They had greater activity on dating-related subreddits and were more likely to write about communication, dating, and texting, as opposed to using wedding-related subreddits or writing about mental health or marriage. In terms of post style, repeat advice-seekers on r/relationships had shorter first posts and used more conjunctions, but user return was negatively related to preposition use. First-person singular pronoun usage was positively related to user return.

4.5 Second analysis: repeat vs non-repeat advice-seekers (split by gender)

The second set of analyses addressed the same basic question of what distinguished the users who returned for advice from the users who did not, but split the $n=17896$ sample by gender, resulting in two separate models.

4.5.1 Justification for creating separate models by gender. When examining the results of the first model by gender, it was found that the model was better at picking out female repeat-advice seekers versus male repeat-advice seekers (68.6% correct vs 47.4% correct) and male non-repeat advice-seekers versus female non-repeat advice seekers (64.2% correct vs 51.3% correct). Based on the statistical tests and model, female users were clearly simply more likely to be repeat advice-seekers than male users – so it would make sense for the model to more often predict female users to be repeat advice-seekers.

However, given that past research has found gendered ways in which users behave online and write (Chapter 2), issues may arise when mixing both genders in one model. Perhaps due to the greater number of female users in the repeat advice-seeker group (55.62% female users versus 43.46% male users) the model may be biased towards using female-gendered cues to predict for repeat advice-seekers. Male repeat advice-seekers – who may behave

differently from female advice-seekers – would thus be less likely to be predicted correctly for.

To ascertain if there is indeed gendered behaviour within the r/relationships dataset, the four groups (i.e., male non-repeat advice-seekers, male repeat advice-seekers, female non-repeat advice-seekers, female repeat advice-seekers) were statistically compared. Only data used in training the model in the first set of analyses was used for this task.

A chi-square test for association found a significant association between the four groups and whether the user's r/relationships post in the 70520-dataset was their first activity on Reddit, $X^2(3) = 76.02$, Cramer's $V = .073$, $p < .001$. The r/relationships post in the 70520-dataset being the user's first activity on Reddit was the case for higher percentages of female non-repeat (29.5%) and repeat advice-seekers (29.5%), versus male non-repeat (23%) and repeat advice-seekers (23.5%). Between male repeat advice-seekers and male non-repeat advice-seekers only, no significant association was found between whether the user was a repeat/non-repeat advice-seeker and whether the user's r/relationships post in the 70520-dataset was their first activity on Reddit, $X^2(1) = .318$, $p = .573$. A similar comparison for female users showed no significant association as well, $X^2(1) = 2.442$, $p = .118$.

Kruskal-Wallis tests with Holm-Bonferroni correction ($\alpha = .05$) were performed on all continuous variables used in training the first model (66 variables) in SPSS 20. Again, non-parametric tests were performed due to the skewness of multiple variables. The full table of all significant results after correction can be found in Appendix E. In summary, most variables showed evidence of gendered behaviour (i.e., female repeat advice-seekers were more similar to female non-repeat advice-seekers than male repeat advice-seekers, and vice versa). This provided evidence for separating the genders when conducting the analyses.

4.5.2 Male user-only prediction model. The male user model involved 9051 users, 11.6% who were repeat advice-seekers. 80% of the data was used

for the training set, and 20% for the testing set. Group proportions were kept roughly equal in training and testing sets. Model building was done only on the training set. The testing set was held-out to evaluate the performance of the trained model. Again, repeat advice-seekers were the positive class (1) and non-repeat advice-seekers the negative class (0).

Model training and selection. L1-penalised logistic regression models were created (Pedregosa et al., 2011). The sets of variables were again evaluated individually, then incrementally based on their individual predictive utility. Performance was evaluated at each step using 5-cross-validation folds on the training set. The best performing model was selected, trained on the full training dataset, then tested on the held-out testing set. Refer to Appendix F for full details on model training.

Results from best performing model. The best performing model was a combination of user age, account information, activity information, post LDA topics, post stylistic information and post sentiment information (full performance metrics of model on testing set in Appendix F). The model's performance on the testing set is displayed in a normalised confusion matrix (Table 5).

Table 5. *Normalised confusion matrix for male-only model from second set of analyses.*

	Predicted as non-repeat advice-seeker	Predicted as repeat advice-seeker
Actual non-repeat advice-seeker	0.58	0.42
Actual repeat advice-seeker	0.40	0.60

60% of the male repeat advice-seekers were correctly identified as such, and 58% of male non-repeat advice-seekers correctly identified (versus 47.4% and 64.2% respectively from the model built in the first set of analyses). Comparing against a dummy classifier which outputs predictions based on the class proportions of the training set (Pedregosa et al., 2011), the dummy

classifier guessed 87% of the non-repeat advice-seekers correctly, but only got 15% of the repeat advice-seekers correct.

To get a sense of the variables related to predicting if a male user was a repeat/non-repeat advice-seeker, the non-zero coefficients of the final model were examined. The full table of coefficients can be found in Appendix F. Table 6 shows the variables positively and negatively related to male repeat advice-seekers.

Table 6. *Variables positively/negatively related to male repeat advice-seekers.*

Positively related to male repeat advice-seekers	Negatively related to male repeat advice-seekers
Total count of account activity in preceding year	Age of user
Prior comments on r/relationships in preceding year	Account age
Activity on dating-related subreddits in preceding year	Drinking/drugs/religion (post LDA)
Activity on advice/help-related subreddits in preceding year	Finances (post LDA)
Activity on mental health-related subreddits in preceding year	Jobs (post LDA)
Activity on anxiety-related subreddits in preceding year	Marital (post LDA)
Dating (post LDA)	Length of post
Social media/trust issues (post LDA)	Usage of articles
Texting (post LDA)	Usage of prepositions
Usage of first-person singular pronouns (e.g. 'I')	
Usage of third-person singular pronouns (e.g. 'she')	
Usage of conjunctions	
Usage of anxiety words	

Summary of results. Results from the model for male-only users were largely similar to those from the first model. They were younger and with

younger but more active Reddit accounts, and were already making greater use of Reddit as a source for help. They were also participating in discussions of the problems of others on r/relationships prior to their first advice-seeking attempt on the forum. They were similarly focused on dating instead of marriage or jobs/finances. Longer posts and usage of articles and prepositions were negatively related to user return. Conversely, conjunctions, first-person singular, and third-person singular pronouns were positively related to user return.

Texting and social media use and the issues that come with its use appeared to be focal in male repeat advice-seekers. A notable difference in the results between the male-only model and the first model was the preponderance of anxiety. Male user return was positively related not only to anxiety-related subreddit use, but also use of anxiety-related words in their writing.

4.5.3 Female user-only prediction model. The female user model involved 8845 users, 15% who were repeat advice-seekers. 80% of the data was used for the training set, and 20% for the testing set. Group proportions were kept roughly equal in training and testing sets. Model building was done only on the training set. The testing set was held-out to evaluate the performance of the trained model. Again, repeat advice-seekers were the positive class (1) and non-repeat advice-seekers the negative class (0).

Model training and selection. L1-penalised logistic regression models were created (Pedregosa et al., 2011). The sets of variables were again evaluated individually, then incrementally based on their individual predictive utility. Performance was evaluated at each step using 5-cross-validation folds on the training set. The best performing model was selected, trained on the full training dataset, then tested on the held-out testing set. Refer to Appendix G for full details on model training.

Results from best performing model. The best performing model was a combination of user age, account information, activity information, post LDA topics, and cognitive process word usage in received comments (full

performance metrics of model on testing set in Appendix G). The model's performance on the testing set is displayed in a normalised confusion matrix (Table 7).

Table 7. *Normalised confusion matrix for female-only model from second set of analyses.*

	Predicted as non-repeat advice-seeker	Predicted as repeat advice-seeker
Actual non-repeat advice-seeker	0.56	0.44
Actual repeat advice-seeker	0.35	0.65

65% of the female repeat advice-seekers were correctly identified as such, and 56% of female non-repeat advice-seekers correctly identified (versus 68.6% and 51.3% respectively from the model built in the first set of analyses). Comparing against a dummy classifier which outputs predictions based on the class proportions of the training set (Pedregosa et al., 2011), the dummy classifier guessed 83% of the non-repeat advice-seekers correctly, but only got 17% of the repeat advice-seekers correct.

To get a sense of the variables related to predicting if a female user is a repeat/non-repeat advice-seeker, the non-zero coefficients of the final model were examined. The full table of coefficients can be found in Appendix G. Table 8 shows the variables positively and negatively related to female repeat advice-seekers.

Table 8. *Variables positively/negatively related to female repeat advice-seekers.*

Positively related to female repeat advice-seekers	Negatively related to female repeat advice-seekers
Total count of account activity in preceding year	Age of user
Prior comments on r/relationships in preceding year	Account age
Activity on dating-related subreddits in preceding year	Activity on sexual orientation-related subreddits in preceding year
Activity on relationship-related subreddits in preceding year	Marital (post LDA)

Activity on advice/help-related subreddits in preceding year	Mental health (post LDA)
Activity on anxiety-related subreddits in preceding year	Trips/occasions (post LDA)
Activity on parental figure-related subreddits in preceding year	
Communication (post LDA)	
Sexual (post LDA)	
Texting (post LDA)	
Usage of causation words in received comments	

Summary of results. Again, female repeat advice-seekers tended to be younger with younger and more active Reddit accounts. Prior usage of Reddit as a source of help was also positively related to female user return. They were also participating in discussions of the problems of others on r/relationships prior to their first advice-seeking attempt on the forum.

However, while usage of dating-related subreddits was positively related to female user return and writing about marriage negatively related, writing about dating was not a non-zero coefficient for female users. Writing about texting was positively related to female user return as with male user return, but so were writing about communication and sex (which were zero coefficients in the male user model). Also, post style was not useful in distinguishing between female repeat and non-repeat advice-seekers unlike in the first and male user models. Surprisingly, usage of causation words by commenters on the female user's post was positively related to female user return.

4.5.4 Comments on differences between male/female user models.

The separated male and female user models did not show a huge departure from the model from the first analysis. Repeat advice-seekers from both genders were characterised as younger in age with younger and more active Reddit accounts. Their activity history showed signs of using Reddit as a source of help – in general and for dating. They were both less focused on longer-term

relationship concerns like marriage. However, there were some differences. Against other male users, male user return was positively predicted by usage of anxiety words in the first advice-seeking post on r/relationships, and by certain ways of writing (i.e., conjunctions, first-person and third-person singular pronouns). Female user return was not predicted by any of these – instead, what had additional utility in predicting returning users from a female-only sample was the usage of causation words in received comments.

For both gender-separated models, though performance overall did not see a drastic boost, their performance in identifying both repeat and non-repeat advice-seekers was more balanced than in the first model. The relatively better ability of the separate models to pick out male repeat advice-seekers and female non-repeat advice-seekers, along with the differences in final predictive variables, highlight a need to consider gender when evaluating such online behaviour.

4.6 Third analysis: accounting for advice-giving behaviour

The third set of analyses built upon the basic question of what distinguished the users who returned for advice from the users who did not by adding the consideration of advice-giving behaviour. In other words, four – instead of two – groups of users were examined: (1) users who came back to give *and* seek advice, (2) users who came back to seek but not give advice, (3) users who came back to give but not seek advice, and (4) users who did not come back to give or seek advice.

4.6.1 Justification for including advice-giving behaviour. Though some users may not come back to seek for more advice after their first post on the forum, they may return to give advice. This behaviour is arguably different from users who do not return in any form – however, in the first two sets of analyses, these users were undistinguished from the other non-repeat advice-seekers.

4.6.2 Creation of user groups. A user was considered to have given advice to another user if they wrote a direct/top-level comment to another user

in the three months following their own first post in the 70520-dataset. The procedure for creating the sample for this analysis, building upon the $n=17896$ sample used in the earlier analyses, was:

- (1) Addition of another 1949 users. These 1949 users had account activity prior to their first post in the 70520-dataset, but were removed from the earlier analyses as they displayed no other behaviour after their first post beyond commenting on r/relationships. These users were relevant for this analysis which included advice-givers.
- (2) Removal of 5328 users who posted in January 2018 or later. This was for the three-month allowance for returning to give advice; users who posted in January 2018 or later needed an observation period that spanned past the 70520-dataset's cut-off of March 2018.

This left 14155 users for analysis. The breakdown of the sample, combined/separated by gender, is shown in Table 9. Interestingly, though proportions of male and female users who returned just to seek advice are comparable (8.29% and 8.95% respectively), a relatively greater proportion of female users returned only to give advice (5.99% of male users vs 11.01% of female users) and returned to both seek and give advice (2.44% of male users vs 4.74% of female users).

Table 9. Breakdown of $n=14155$ user activity in 3 months from first post in dataset.

	Returned to seek advice + Returned to give advice	Did not return to seek advice + Returned to give advice	Returned to seek advice + Did not return to give advice	Did not return to seek advice + Did not return to give advice
Male users ($n=7242$)	177 (2.44%)	434 (5.99%)	600 (8.29%)	6031 (83.28%)
Female users ($n=6913$)	328 (4.74%)	761 (11.01%)	619 (8.95%)	5205 (75.29%)
Total ($n=14155$)	505 (3.57%)	1195 (8.40%)	1219 (8.61%)	11236 (79.38%)

4.6.3 Prediction task. The limited sizes of some groups if broken down by gender (e.g. 177 male users returning to seek and give advice) would have made training difficult. Thus, the current analysis did not separate the users by gender. Again, 80% of the $n=14155$ sample was used for training and 20% for testing. Group proportions were kept roughly equal in training and testing sets. Model building was done only on the training set; the testing set was held-out for evaluation of the trained model.

Model training and selection. L2-penalised multinomial logistic regression models were created (Pedregosa et al., 2011). Regularisation with L2, as with L1, was used to avoid overfitting due to the large number of variables. However, instead of pushing some coefficients to exactly zero (i.e. L1), L2 shrinks coefficients towards zero. The sets of variables were again evaluated individually, then incrementally based on their individual predictive utility. Performance was evaluated at each step using 5-cross-validation folds on the training set. The best performing model was selected, trained on the full training dataset, then tested on the held-out testing set. Refer to Appendix H for full details on model training.

Results from best performing model. The best performing model was a combination of user information, account information, activity information, post LDA topics, post stylistic information and cognitive process word usage in the post (full performance metrics of model on testing set in Appendix H). For clarity, the model's performance on the testing set is displayed in a normalised confusion matrix (Table 10).

Table 10. *Normalised confusion matrix for multinomial model from third set of analyses.*

	Predicted as full non-return	Predicted as advice-giving	Predicted as advice-seeking	Predicted as full return
Actual full non-return	0.45	0.13	0.30	0.12
Actual advice-giving	0.20	0.41	0.19	0.20
Actual advice-seeking	0.29	0.14	0.39	0.18
Actual full return	0.20	0.26	0.31	0.23

Full non-return: did not return to seek/give advice

Full return: returned to seek and give advice

45% of users who never returned to r/relationships after their first post were correctly identified. 41% of those who returned only to give advice to others after their first post were correctly identified. 39% of those who returned only to seek advice again after their first post were correctly identified. Finally, 23% of those who returned to seek and give advice after their first post were correctly identified. In comparison, a dummy classifier guessing based on class proportions in the training dataset guessed 79% of the majority class (full non-return) correctly, 8% of the advice-giving users correctly, 8% of the advice-seeking users correctly, and 3% of the full-return users correctly.

The coefficients derived from the multinomial logistic regression model can be found in Appendix H. Several variables found from the prior analyses to

be helpful in predicting a repeat advice-seeker held here. Younger user age, higher activity on dating-, relationship-, advice/help-, confession- and anxiety-related subreddits were related to a higher probability of a user being a repeat advice-seeker. Higher focus on communication, dating, texting in the first post and shorter posts were also related to a higher probability of the user being a repeat advice-seeker.

However, with the more nuanced breakdown in user behaviour, features like count of comments on r/relationships in the year preceding the first post and total account activity – which were predictive of repeat advice-seekers in prior analyses – were now more likely to be predictive of returning advice-givers and full returning users (sought *and* gave advice).

4.7 Fourth analysis: ‘veteran’ repeat advice-seekers vs non-repeat advice-seekers

The last set of analyses returned to and extended upon the basic question of what distinguished the repeat advice-seekers from non-repeat advice-seekers, by using a sample of users who have sought for advice on r/relationships prior to the time period covered by the 70520-dataset. As no prediction task was formulated for this analysis, the full dataset of $n=12217$ users (788 ‘veteran’ repeat advice-seekers, 11429 non-repeat advice-seekers; 6.45% ‘veteran’ repeat advice-seekers) was used for the statistical analyses. The selection of these users was covered in the earlier section on user selection (4.2 User selection).

4.7.1 Justification for examining ‘veteran’ repeat advice-seekers.

The purpose of this analysis was to check if the earlier-identified features distinguishing between the first-time returnees and non-returnees were consistent and could be observed even in a completely separate sample of users who, based on their past activity, were even more definitively advice-seeking returnees.

4.7.2 Statistical tests. Based on a chi-square test of independence, there was a significant association between gender and whether the user returned for

advice, $X^2(1) = 75.83$, $\Phi = .079$, $p < .001$. Mann-Whitney U-tests with Holm-Bonferroni correction ($\alpha = .05$) for multiple comparisons (147 comparisons) for the continuous variables were performed. Non-parametric tests were performed due to the skewness of multiple variables. Only significant results kept after the Holm-Bonferroni correction are reported in Table 11.

Table 11. *Significant results after Holm-Bonferroni correction ('veteran' repeat vs non-repeat advice-seekers).*

Variable	Group higher in mean ranks	Cohen's d
Account age at time of first r/relationships post in 70520-dataset	Non-repeat advice-seekers	.144***
Post LDA topic: Communication	Repeat advice-seekers	.088***
Readability of post	Repeat advice-seekers	.069***
Length of post	Non-repeat advice-seekers	.108***
Seconds between post and first comment	Non-repeat advice-seekers	.082***
Total number of comments received on post	Repeat advice-seekers	.089***
Usage of first-person singular pronouns	Repeat advice-seekers	.087***
Usage of first-person plural pronouns	Non-repeat advice-seekers	.087***
Usage of prepositions	Non-repeat advice-seekers	.12***
Usage of conjunctions	Repeat advice-seekers	.072***
Comments on r/relationships for one year prior to first post in 70520-dataset	Repeat advice-seekers	.604***
Activity on Reddit for one year prior to first post in 70520-dataset	Repeat advice-seekers	.238***
Activity on validation-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.025***
Activity on dating-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.131***

Activity on relationship-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.109***
Activity on sex-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.089***
Activity on advice/help subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.114***
Activity on confession-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.043***
Activity on mental health-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.022***
Activity on anxiety-related subreddits for one year prior to first post in 70520-dataset	Repeat advice-seekers	.061***

*** $p < .001$

A number of findings in this sample replicated significant findings from the first set of analyses, namely:

- (1) Repeat advice-seekers had younger Reddit accounts.
- (2) The posts of repeat advice-seekers generally covered more about communication than non-repeat advice-seekers.
- (3) The posts of repeat advice-seekers were typically shorter versus non-repeat advice-seekers.
- (4) Repeat advice-seekers generally used more conjunctions than non-repeat advice-seekers.
- (5) Repeat advice-seekers were more active on dating-related, relationship-related, advice/help-related, and anxiety-related subreddits versus non-repeat advice-seekers in the year prior to their first post in the 70520-dataset.
- (6) Repeat advice-seekers generally posted more comments on r/relationships in the year prior to their first post in the 70520-dataset.

Some variables in this set of analyses did not pass the correction, but followed the direction of some significant findings from the first set of analyses:

- (1) Repeat advice-seekers were generally younger in age versus non-repeat advice-seekers ($p=.0008$, $d=.06$).
- (2) Repeat advice-seekers typically wrote more about dating ($p=.007$, $d=.049$) and texting/calling ($p=.051$, $d=.035$) versus non-repeat advice-seekers; comments received by repeat advice-seekers tended to write more on communication versus those received by non-repeat advice-seekers ($p=.014$, $d=.045$).
- (3) Repeat advice-seekers were writing less about mental health ($p=.049$, $d=.036$) versus non-repeat advice-seekers; the comments received by returnees also were less about mental health (therapy) ($p=.002$, $d=.057$).

On the other hand, some variables from the first set of analyses did not pass the correction, but followed the direction of some significant findings from this set of analyses:

- (1) The posts of repeat advice-seekers were generally easier to read ($p=.001$, $d=.055$).
- (2) The posts of repeat advice-seekers generally used more first-person singular pronouns ($p=.001$, $d=.056$) and less prepositions ($p=.001$, $d=.057$).
- (3) The posts of repeat advice-seekers typically received more comments than non-repeat advice-seekers ($p=.042$, $d=.034$).
- (4) Repeat advice-seekers generally participated more in confession-related subreddits ($p=.024$, $d=.011$).

However, several significant findings from this analysis were not observed in the first set of analyses:

- (1) The posts of ‘veteran’ repeat advice-seekers typically used less first-person plural pronouns. This was not evident in the first analysis ($p=.540$).
- (2) The posts of ‘veteran’ repeat advice-seekers typically received their first comment faster than the posts of non-repeat advice-seekers. This was not observed in the first analysis ($p=.090$).
- (3) Though ‘veteran’ repeat advice-seekers may be newer to Reddit versus non-repeat advice-seekers, they generally showed greater activity (posting and commenting) on Reddit for the one year prior to their first post in the 70520-dataset. This was not apparent in the statistical tests of the first analysis ($p=.853$), though it was a non-zero coefficient in the prediction models.
- (4) ‘Veteran’ repeat advice-seekers generally participated more in validation-related, sex-related, and mental health-related subreddits in the one year prior to their first post in the 70520-dataset versus non-repeat advice-seekers. This was not evident in the first set of analyses, with little difference in mean ranks between non-repeat and repeat advice-seeker groups (validation: $p=.351$; sex: $p=.116$; mental health: $p=.286$).

CHAPTER V: DISCUSSION

5.1 Overview of chapter

The study explored how (1) user-level differences and (2) differences in forum community reception may be associated with a user deciding to seek advice again on the same advice forum, r/relationships. Recall that repeat advice-seekers were r/relationships posters who sought for advice within the next three months following their first post on r/relationships. Key user-level cues predictive of repeat advice-seekers were uncovered in the first set of analyses (e.g. repeat advice-seekers typically had greater past activity on advice/help-related subreddits and greater use of conjunctions in their first advice-seeking posts on r/relationships). To refine the findings of the first set of analyses, the second set of analyses then separated the users by gender and the third set of analyses accounted for the complementary behaviour of advice-giving. The final set of analyses with ‘veteran’ repeat advice-seekers (repeat advice-seekers whose first post on r/relationships predated the 70520-dataset) provided affirmation for the findings from the first set of analyses, while also highlighting a number of key differences.

The models created for the first to third sets of analyses all performed better than baseline though they did not achieve very high accuracies. Still, it must be emphasised that the goal of the project was not to establish deployment-ready models of the phenomenon; it is emphasised that the model as is should not be applied to real-world purposes. Instead, the focus of the project was to gain some initial and better understanding of the factors related to repeat online advice-seeking, with the predictive models built to ascertain that the relationships found were not wholly spurious (Adjerid & Kelley, 2018). Together, the statistical tests and the logistic regression models showed that past user activity on certain types of forums and the content and style of a user’s first advice-seeking post can provide valuable information on whether the user will return to the forum for more advice. These basic observable cues

together were able to predict – better than chance – whether the user returned for more advice, even without knowledge of the user’s activity after their first post and without external information of the user’s personality and extenuating life circumstances.

This final chapter reviews the results of Chapter 4 and concludes with a discussion of the implications and limitations of the study, before wrapping up with potential directions for future research.

5.2 Factors related to repeat advice-seeking on r/relationships

5.2.1 Demographic and account characteristics. Repeat advice-seekers were more likely to be female, younger in age, and with younger and more active accounts versus non-repeat advice-seekers. These characteristics corroborate with the literature – a large survey of American online support seekers found that they were more likely to be female (DeAndrea & Anthony, 2013). Younger adults typically participate more in the online sphere (Booth, Lin & Wei, 2018) and are more inclined toward interacting with online strangers versus older adults (Thayer & Ray, 2006). More active users may also simply be more comfortable seeking for advice on personal issues online – university students who spent more time online were also found to be more unreserved and disclosed more personal matters online versus their peers (Niemz, Griffiths & Banyard, 2005).

5.2.2 Past activity on certain types of subreddits. Compared to non-repeat advice-seekers, repeat advice-seekers typically posted and/or commented more on advice/help-, anxiety-, relationship-, and dating-related subreddits in the year prior to their first post on r/relationships. In other words, repeat advice-seekers were already using Reddit as a source of help, particularly for relational issues. r/relationships may have thus appeared as an additional help resource to these users. Given Reddit’s large pseudonymous environment, other users are not likely to be part of a user’s current primary offline social network. This could be attractive to users who are afraid of their disclosures spreading around their offline networks (Ruble, 2011) and losing face (Walther & Boyd, 2002),

especially since their queries focus on offline relational issues that may involve those very individuals in their network. The usage of anxiety-related subreddits (r/HealthAnxiety, r/Anxiety, r/AskDocs) being telling of repeat advice-seekers matches past literature – online support group usage has been found to be related to health anxiety (Lagoe & Atkin, 2015). A study of Twitter users with anxiety similarly found that increased levels of anxiety were related to subsequent heightened social interaction with weak ties (Dutta, Ma & De Choudhury, 2018). Even in children, feelings of loneliness and social anxiety have been related to the likelihood of them reaching out to unfamiliar people via instant-messaging (Gross, Juvonen & Gable, 2002).

5.2.3 More past comments contributed to r/relationships. Repeat advice-seekers generally contributed more comments on the r/relationships posts of other users in the year preceding their own first post on r/relationships. Being more active on the subreddit, they may have felt that they were part of a community of individuals struggling with relational issues, thus increasing the likelihood of them to return for more advice. This conjecture however cannot be conclusively determined from the current data and would benefit from a user survey.

5.2.4 Content of first r/relationships post. The first posts of repeat advice-seekers had a greater focus on dating, texting/calling, and communication versus non-repeat advice-seekers. On the other hand, they had a weaker focus on topics like mental health and marriage. This focus matched the past subreddit activity focus of repeat advice-seekers. It may also be reflective of the focus of the r/relationships subreddit. For online advice specifically, much of it is delivered indirectly via self-disclosure of the advice giver's personal experiences (White & Dorman, 2001; Harrison & Barlow, 2009). Thus, it could be that more r/relationships users can relate to and share their experiences on more general topics like dating and communication issues. Specific topics like marriage and mental health may on the other hand receive less relevant and/or helpful feedback, perhaps then driving users with these

issues towards more topically-specific subreddits (e.g. r/depression, r/DeadBedrooms) – this can be investigated in future work. In addition, topics like communication and texting may also be more suited for receiving frequent online advice on. Advice typically endorses some specific form of action to follow up on (Brooks et al., 2015). Applying the received advice on one's texting expectations would likely be comparatively easier and faster than trying it out on a marital issue – which could also explain why users who return to r/relationships for advice again within three months typically focus on these issues.

5.2.5 Style of first r/relationships post. The first posts of repeat advice-seekers typically were shorter in length, used more conjunctions and first-person singular pronouns and used less prepositions versus those of non-repeat advice-seekers. Usage of the first-person singular pronoun has been strongly linked to self-focus and self-disclosure (De Choudhury & De, 2014; Wang et al., 2015; De Choudhury et al., 2016; Malloch & Taylor, 2018). Conjunctions – especially simple ones like *and* – commonly occur in texts created under time restrictions (e.g. everyday speech) (Biber, 1988). Together with first-person singular pronouns, conjunctions may be indicative of dynamic texts (e.g. personal stories) (Pennebaker, et al., 2014). Prepositions in contrast are indicative of categorical language seen in formal (Pennebaker et al., 2014) and academic writing (Biber, 1988). These stylistic features indicate that the posts of repeat advice-seekers were generally more personal, self-disclosing and informal than those of non-repeat advice-seekers; they are in fact rather similar to the language style of personal letters (Biber, 1988).

5.3 Going further: considering gender and advice-giving

When accounting for the complementary behaviour of advice-giving, it was observed that several features from the base model (e.g. a focus on dating in the user's first post and past activities, usage of anxiety-related subreddits) were still predictive of users who returned solely for advice. However, other features like the count of prior comments on r/relationships and account activity

in the preceding year were more predictive of users who returned solely to give advice/returned to seek *and* give advice. Future studies thus may want to account for these more nuanced behaviours in the study of online advice-seeking; this study indicates that it may be feasible and worthwhile to further pursue this line of inquiry, especially given that the data is relatively easy to collect and is more naturalistic than asking users to fill self-report surveys.

The current study also draws to attention the importance of considering gender even when working with a pseudonymous environment with no pictorial cues (e.g. Facebook profile pictures). Prior research has highlighted the gendered cues present in online communication (Spottswood, Walther, Holmstrom & Ellison, 2013; De Choudhury et al., 2017). Likewise, for most variables used in the model in the first analysis, there was evidence of gendered behaviour – female repeat advice-seekers were more similar to female non-repeat advice-seekers than they were to male repeat advice-seekers, and likewise for male repeat advice-seekers and male non-repeat advice-seekers. Separately constructing models for male and female users provided better performance and highlighted some differences in selected variables (e.g. usage of anxiety words is predictive for male repeat advice-seekers but not for female repeat advice-seekers).

Furthermore while Reddit has traditionally been viewed as masculine in nature (Massanari, 2016), in the current study, it was found that proportionately more female users versus male users returned to give advice/returned to give *and* seek advice, though the proportions of female and male users who returned to seek advice (but not give advice) were similar. In a telephone survey of 370 Americans, it was found that romantic regrets were more common amongst women than men (Morrison & Roese, 2011). Thus, it could be that the observed gender differences in advice-seeking/giving behaviours were due to the content of this specific subreddit (relational issues). However, it may be worthwhile to verify if they replicate across other help-centred subreddits with different content focuses – though the task would be difficult, given that few subreddits

stipulate self-report of gender like r/relationships. Nevertheless, this study has highlighted the need to factor in how personal offline characteristics like gender may still imprint themselves onto online data, and the possible biases that may arise if analyses and models are constructed without thorough consideration of them.

5.4 Comparison with the ‘veteran’ sample

The final set of analyses compared ‘veteran’ repeat advice-seekers against non-repeat advice-seekers. Recall that ‘veteran’ repeat advice-seekers were advice-seekers who had posted for advice at least once in the three months prior to the 70520-dataset, and then again in the three months following their first post in the 70520-dataset.

5.4.1 Similarities. Several similarities between this comparison and the comparison in the first analysis (repeat advice-seekers versus non-repeat advice-seekers) were noted. Repeat advice-seekers from both analyses tended to be younger in age with younger and more active accounts, and had commented on r/relationships more than non-repeat advice-seekers. Notably, they were also active on dating-, relationship-, advice/help- and anxiety-related subreddits prior to their first post in the 70520-dataset. Writing-wise, both sets of repeat advice-seekers generally wrote shorter posts, used more conjunctions and first-person singular pronouns but less prepositions than non-repeat advice-seekers. Both sets of repeat advice-seekers also typically focused more on communication, dating, and texting/calling in their posts than non-repeat advice-seekers. In other words, there are certain behavioural cues that may be able to consistently distinguish users who keep returning for more advice on r/relationships versus those who do not. The practical implications of this will be discussed later (see 5.5 Implications).

5.4.2 Differences. However, when comparing the ‘veteran’ repeat advice-seekers to non-repeat advice-seekers, some features were found that were not apparent from the first set of analyses.

Usage of first-person plural pronouns. ‘Veteran’ repeat advice-seekers used less first-person plural pronouns in their first post in the 70520-dataset versus non-repeat advice-seekers. Paired with the relatively higher usage of first-person singular pronouns versus non-repeat advice-seekers, this pattern of pronoun usage is similar to that observed in the mental health literature. Specifically, writings of suicidal individuals have been found to use less first-person plural pronouns but more first-person singular pronouns (Chung & Pennebaker, 2007; Kumar, Dredze, Coppersmith & De Choudhury, 2015). This was also observed in the postings of Redditors on anxiety subreddits (Shen & Rudzicz, 2017) and mental health subreddits (Pavalanathan & De Choudhury, 2015; Li, et al., 2018). The higher first-person singular pronoun usage demonstrates greater inward self-focus (Tausczik & Pennebaker, 2010) and the lowered first-person plural pronoun usage suggests social isolation and a lack of attention to others (Chung & Pennebaker, 2007). It would be premature to conclude that the ‘veteran’ repeat advice-seekers are/were suffering from mental health disorders. However, given the link between function word usage and well-being (Kornfield et al., 2018), their pronoun usage indicates that ‘veteran’ repeat advice-seekers may be more vulnerable.

Time to first comment. ‘Veteran’ repeat advice-seekers typically received the first comment to their post faster than non-repeat advice-seekers. This could be related to the way their posts were written (e.g. appearing to need help more urgently) (Balani & De Choudhury, 2015; Kumar et al., 2015). The advice-givers in the community may also ‘recognise’ the user given their constant reappearances on r/relationships and hence reply to them faster. In any case, given that faster replies demonstrate the readiness of the other users in advising (Balani & De Choudhury, 2015), this group of advice-seekers may be encouraged to further rely on r/relationships as they may perceive r/relationships as providing higher quality support (Ledbetter, 2009) and/or being of high availability (Hofmann, Lei & Grant, 2009).

Past activity on other types of subreddits. Compared to non-repeat advice-seekers, repeat advice-seekers made greater use of validation-, sex-, and mental health-related subreddits prior to their first post in the 70520-dataset. This was not observed in the repeat advice-seekers from the first analysis. A number of subreddits under these groups go beyond textual sharing and involve users revealing their faces/bodies and/or commenting on the pictures of other users (e.g. r/Rateme, r/amiugly, r/gonewild), suggesting that ‘veteran’ repeat advice-seekers as a whole may be even more open to others online. Taken together with this group’s pattern of pronoun usage, these behavioural cues again may suggest possible psychological vulnerabilities that can lead to an unhealthy reliance on the Internet (Lee & Stapinski, 2012). Furthermore, their posting behaviour may expose them to risks from other users (Shelton et al., 2015); for example, if they have posted explicit media of themselves on r/gonewild on the same account they use to seek for advice on r/relationships. The current study is limited in what it can conclude with regards to the well-being of these users, but it would be interesting in future to track the development of such user behaviour and writing over a much longer time-frame as these users interact more with the advice forum and the rest of Reddit.

5.5 Implications

The current study provides a first exploratory look at the factors related to online advice-seeking, a phenomenon which despite its size, remains relatively understudied (MacGeorge et al., 2016a; Rains & Wright, 2016). By utilising rich records of naturalistic human behaviour available online, the current study examined a diverse range of real-life stressors and behaviours otherwise difficult to obtain and simulate in controlled laboratory studies (Shaw & Gant, 2002; Adjerid & Kelley, 2018; Collison et al., 2018). In a representative survey of 370 Americans, regrets involving romance and family were the most commonly quoted (Morrison & Roese, 2011). The current study thus addresses online advice-seeking in an important and fundamentally human context – relationships.

It was found that user gender and age, certain stylistic and content cues from a user's first advice-seeking post and usage of a number of subreddit types were related to the user returning to seek for more advice on the forum. These cues are easily retrievable and, as covered earlier, corroborate with existing literature. The study also showed that even with these relatively simplistic and limited cues, advice-seeking can be predicted for better than chance, even without knowledge of important psychological traits like personality and subjective variables like the user's evaluation of the advice (see 5.6 Limitations, for further discussion). Though the current models do not achieve high predictive accuracy, this study does provide some evidence that predicting user reliance on the forum – a complex and multifaceted phenomenon – may be a viable undertaking, especially with further refinement of the features used. With a wider range of features and a stronger emphasis on features for past user activity given their predictive importance, future models may achieve much better results.

Understanding the factors related to user return to an advice forum – and consequently perhaps being able to predict it reliably – is an important task. A concern that frequently arises in the literature is that some individuals who display a penchant for online interaction may be using it as a way to compensate for the inadequacies of their offline social lives (Livingstone & Helsper, 2007; Mazzoni, Baiocco, Cannata & Dimas, 2016). Overreliance in such cases may worsen pre-existing distress (Lawlor & Kirakowski, 2014). The current study does not provide enough evidence to confirm whether repeat advice-seekers fall under this problematic category of users and was not designed to address whether repeat advice-seeking is a negative behaviour. What the results from the 'veteran' repeat advice-seeker sample (i.e., pronoun usage similar to those found in mental health samples, usage of face/body-revealing and mental health subreddits along with advice-, relationship- and anxiety-related subreddits) do indicate at this stage is that more attention needs to be paid to studying users who consistently return, as they may be a

potentially vulnerable group. Furthermore, given that a recent study has noted that users who specifically seek for health-related information on Reddit were more likely to perform the suggestions (Record, Silberman, Santiago & Ham, 2018), it may be useful just to be able to identify users who are likely to return and understand the issues that they require advice on. The responses of the forum to such users and topics can then perhaps be more closely moderated, given the possible propensity of this group of users to apply received advice to their real-life relationships.

5.6 Limitations

The main limitation in the current study concerns the handling of the comments. The lack of significance of the comment features does not necessarily mean that comments are completely unimportant to users. Receiving a response has been found to be an important factor in whether a user will return to an online group (Arguello et al., 2006; Joyce & Kraut, 2006), but in the current study, posts that received no comments were excluded for the purpose of focusing on *how* the content/style of the received comments may be related to repeat advice-seeking. Furthermore, though the comment features were created based on past theory, it is possible that they were simply too low-level. A future project might benefit from more high-level features (e.g. degree of emotional support present in the advice (MacGeorge et al., 2016b)). Such features can be created by manual annotation by independent coders to create a large enough dataset, which can then in turn be used to train and refine a classifier using low-level features similar to those used in this project to help automate the labelling of the remaining comments. In addition, for the current study, all top-level comments in each post were weighted equally in importance despite the fact that they likely did not score similarly and were delivered at different times. Future studies could look at treating top-scoring or early comments differently; the first few responses may be more impactful to the user (Malloch & Taylor, 2018).

There are also limitations on what can be gleaned from such behavioural data. As the current study required the examined posts to have identifiable usernames to track user history, deleted accounts had to be dropped – there was no manner through which a deleted user’s history can be retrieved, even though some of the deleted users may have posted multiple times on r/relationships. Additionally, the current study (as with similar studies) was unable to capture the number of posts from other users that the user may have read for supplementary secondary advice (Wang et al., 2012). The users may also have received advice via private messages or had developments offline (Soldaini et al., 2018); this information is not found publicly.

User characteristics like personality were also not available. There are many subjective aspects of advice that contribute to the recipient’s evaluation of it that were not possible to capture in this study as well. Notable aspects include how effective the advice would be at resolving the advice-seeker’s problem and whether the advice-seeker would be able to carry out the advice (Feng & Burleson, 2008), along with how polite the advice appears to the advice-seeker (Locher, 2006) and agrees with their pre-existing beliefs (Feng & MacGeorge, 2010; Sillence, 2010).

5.7 Future directions

An immediate follow-up could be to consider the frequency of creating new advice-seeking posts on r/relationships following the advice-seeker’s first r/relationships post. As shown in the analysis of ‘veteran’ repeat advice-seekers, users who show a consistent pattern of returning within a relatively short time-span may be a more vulnerable sample. In addition, given the predictive importance of past posting activity on certain types of subreddits, to observe the effect of posting and the comments on r/relationships more clearly, it could be interesting to control for account history using a method like propensity score matching (e.g. De Choudhury & Kiciman, 2017).

Future studies can also take into account gender in analyses, given the evidence from the current study regarding gendered behaviour – for example,

how users respond differently based on the self-reported gender of the user and how this difference in response may in turn be related to different rates of return. A recent study has indeed found that female-sounding usernames and writing cues are related to more supportive *and* negative comments (Wang & Jurgens, 2018). As many users also report the gender of the other individual(s) involved in their problem, this valuable information can be incorporated too.

The work on examining the user's advice-giving behaviour in tangent with their advice-seeking behaviour following their first r/relationships post can also be expanded upon. The current study found that proportionately more female users than male users were returning to give advice/returning to give *and* seek advice. Previous work has found that individuals reported greater confidence and motivation after giving advice – instead of receiving advice – on a domain they were struggling with in terms of motivation (e.g. losing weight) (Eskreis-Winkler, Fishbach & Duckworth, 2018). A question that arises in the current study's context then is if these female users are giving advice on the same topics that they are seeking help for; thus inadvertently employing advice-giving as another coping mechanism that provides benefits for the handling of their own problems (similar to the motivational increases found in Eskreis-Winkler et al's (2018) studies)? And, if so – why are proportionally more female users than male users engaging in this behaviour?

Future work may also benefit from having a complementary survey study of users of such sites. Besides verifying the findings of this study (e.g. with regards to gender and indications of anxiety), surveys can help assess the relationships between uncaptured user variables like personality and the behaviour of repeated online advice-seeking. Focus group sessions with users of such sites can also help clarify the motivations and intentions involved when users choose to repeatedly seek advice online.

Finally, Reddit – and by extension, r/relationships – comes with its own set of cultural norms and expectations for user self-presentation. Thus, the content and writing style of users studied here may not necessarily carry over to

other sites (e.g. Ask MetaFilter, Quora). Furthermore, the sample of users studied here is relatively young –in the $n=17896$ sample, the mean age was 24.10 with a standard deviation of 5.36 – and so might not be representative of advice forums elsewhere. The shortcomings of using just one site as a model for understanding phenomena have been highlighted (Tufekci, 2014), thus making it imperative to examine how advice-seeking behaviour is performed across other question and answer/advice forums. The diversity of the modern global cyberspace provides a great opportunity to address how behaviours stay constant (or change) across different communities.

REFERENCES

- [reddit change] Selfpost character limit is now 40,000 for all subreddits : changelog. (2018). reddit. Retrieved 22 April 2018, from https://www.reddit.com/r/changelog/comments/39hf9x/reddit_change_selfpost_character_limit_is_now/
- Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and yahoo answers. Proceeding of the 17th international conference on World Wide Web - WWW '08. doi:10.1145/1367497.1367587
- Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, 73(7), 899-917. doi:10.1037/amp0000190
- Ahmed, R. (2017). *Internet Is Furious Over Jealous Wife Who Asked Reddit For Help As She Did Something 'Unforgivable'*. RearFront. Retrieved 22 April 2018, from <https://www.rearfront.com/jealous-wife-did-unforgivable/>
- Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014). How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. In *ICWSM*. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8106/8101>
- Arguello, J., Butler, B. S., Joyce, L., Kraut, R., Ling, K. S., & Wang, X. (2006). Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*. doi:10.1145/1124772.1124916
- Balani, S., & De Choudhury, M. (2015). Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1373-1378). ACM. doi: 10.1145/2702613.2732733
- Barak, A., Boniel-Nissim, M., & Suler, J. (2008). Fostering empowerment in online support groups. *Computers In Human Behavior*, 24(5), 1867-1883. doi:10.1016/j.chb.2008.02.004
- Barak, A., & Gluck-Ofri, O. (2007). Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior*, 10(3), 407-417. doi:10.1089/cpb.2006.9938
- Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10), 1531-1542. doi:10.1177/0956797615594620
- Barbosa, S., Cosley, D., Sharma, A., & Cesar Jr, R. M. (2016). Averaging gone wrong: Using time-aware analyses to better understand behavior. In *Proceedings of the 25th International Conference on World Wide Web*

- (pp. 829-841). International World Wide Web Conferences Steering Committee. doi:10.1145/2872427.2883083
- Bazarova, N. N., Choi, Y. H., Whitlock, J., Cosley, D., & Sosik, V. (2017). Psychological distress and emotional expression on Facebook. *Cyberpsychology, Behavior, and Social Networking*, 20(3), 157-163. doi: 10.1089/cyber.2016.0335
- Barrera, M. (1986). Distinctions between social support concepts, measures, and models. *American Journal Of Community Psychology*, 14(4), 413-445. doi:10.1007/bf00922627
- Biber, D. (1986). On the investigation of spoken/written differences. *Studia Linguistica*, 40(1), 1-21. doi:10.1111/j.1467-9582.1986.tb00759.x
- Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77. doi:10.1145/2133806.2133826
- Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022. doi:10.1162/jmlr.2003.3.4-5.993
- Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. doi:10.1088/1742-5468/2008/10/p10008
- Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2), 127-151. doi:10.1016/j.obhdp.2006.07.001
- Booth, J. M., Lin, Y., & Wei, K. (2018). Neighborhood disadvantage, residents' distress, and online social communication: Harnessing Twitter data to examine neighborhood effects. *Journal of Community Psychology*, 46(7), 829-843. doi:10.1002/jcop.21975
- boyd, D., & Ellison, N. (2010). Social network sites: definition, history, and scholarship. *IEEE Engineering Management Review*, 38(3), 16-31. doi:10.1109/emr.2010.5559139
- Boyd, R. L., & Pennebaker, J. W. (2016). A way with words: Using language for psychological science in the modern era. In C. V. Dimofte, C. P. Haugtvedt, & R. F. Yalch (Eds.), *Consumer psychology in a social media world* (pp. 222-236). New York, NY, US: Routledge/Taylor & Francis Group.
- Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: a new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, 18, 63-68. doi:10.1016/j.cobeha.2017.07.017
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

- Brooks, A. W., Gino, F., & Schweitzer, M. E. (2015). Smart people ask for (my) advice: Seeking advice boosts perceptions of competence. *Management Science*, 61(6), 1421-1435. doi:10.1287/mnsc.2014.2054
- Brunet, P. M., & Schmidt, L. A. (2007). Is shyness context specific? Relation between shyness and online self-disclosure with and without a live webcam in young adults. *Journal of Research in Personality*, 41(4), 938-945. doi:10.1016/j.jrp.2006.09.001
- Caplan, S. E., & Samter, W. (1999). The role of facework in younger and older adults' evaluations of social support messages. *Communication Quarterly*, 47(3), 245-264. doi:10.1080/01463379909385558
- Carlson, C. L. (2016). Predicting emerging adults' implementation of parental advice: Source, situation, relationship, and message characteristics. *Western Journal of Communication*, 80(3), 304-326. doi:10.1080/10570314.2016.1142112
- Carron-Arthur, B., Reynolds, J., Bennett, K., Bennett, A., Cunningham, J. A., & Griffiths, K. M. (2016). Community structure of a mental health internet support group: Modularity in user thread participation. *JMIR Mental Health*, 3(2), e20. doi:10.2196/mental.4961
- Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*. doi:10.1145/2818048.2819973
- Chang, J., & Danescu-Niculescu-Mizil, C. (2019). Trajectories of blocked community members: Redemption, recidivism and departure. *The World Wide Web Conference on - WWW '19*. doi:10.1145/3308558.3313638
- Chen, E., & Wojcik, S. (2016). A practical guide to big data research in psychology. *Psychological Methods*, 21(4), 458-474. doi:10.1037/met0000111
- Chmiel, A. A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., & Holyst, J. A. (2011). Negative emotions boost user activity at BBC forum. *Physica A: Statistical Mechanics and its Applications*, 390(16), 2936-2944. doi:10.1016/j.physa.2011.03.040
- Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions: "On the Internet, Nobody Knows You're a Dog". *Computers in Human Behavior*, 23(6), 3038-3056. doi:10.1016/j.chb.2006.09.001
- Chung, T., & Chen, Y. (2018). Exchanging social support on online teacher groups: Relation to teacher self-efficacy. *Telematics And Informatics*, 35(5), 1542-1552. doi: 10.1016/j.tele.2018.03.022
- Chung, C., & Pennebaker, J. (2007). The Psychological Functions of Function Words. In K. Fiedler (Ed.), *Frontiers of social psychology. Social communication* (pp. 343-359). New York, NY, US: Psychology Press.

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
doi:10.1177/001316446002000104
- Cohen, S., & Wills, T. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, 98(2), 310-357. doi:10.1037//0033-2909.98.2.310
- Collisson, B., Cordoviz, P., de Leon, L. P., Guillen, S., Shier, J., & Xiao, Z. (2018). "Should I Break Up or Make Up?" A Text Analysis of Online Relationship Advice. *North American Journal of Psychology*, 20(2), 301-310. doi: 10.1155/2012/490647
- Colvin, J., Chenoweth, L., Bold, M., & Harding, C. (2004). Caregivers of Older Adults: Advantages and Disadvantages of Internet-based Social Support. *Family Relations*, 53(1), 49-57. doi:10.1111/j.1741-3729.2004.00008.x
- Cotten, S. R., & Gupta, S. S. (2004). Characteristics of online and offline health information seekers and factors that discriminate between them. *Social Science & Medicine*, 59(9), 1795-1806.
doi:10.1016/j.socscimed.2004.02.020
- Coursaris, C., & Liu, M. (2009). An analysis of social support exchanges in online HIV/AIDS self-help groups. *Computers In Human Behavior*, 25(4), 911-918. doi:10.1016/j.chb.2009.03.006
- Cunha, T. O., Weber, I., Haddadi, H., & Pappa, G. L. (2016). The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference* (pp. 99-103). ACM. doi:10.1145/2896338.2897732
- Cunha, T., Weber, I., & Pappa, G. (2017). A warm welcome matters! The link between social feedback and weight loss in /r/loseit. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. doi:10.1145/3041021.3055131
- Cutrona, C., & Suhr, J. (1992). Controllability of Stressful Events and Satisfaction With Spouse Support Behaviors. *Communication Research*, 19(2), 154-174. doi:10.1177/009365092019002002
- Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes*, 112(1), 11-23. doi:10.1016/j.obhdp.2009.11.007
- Danish, Dahiya, Y., & Talukdar, P. (2015). Want Answers? A Reddit Inspired Study on How to Pose Questions. arXiv preprint arXiv:1512.01768. Retrieved from <https://arxiv.org/pdf/1512.01768.pdf>
- Danon, L., Díaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09), P09008-P09008. doi:10.1088/1742-5468/2005/09/p09008

- Davis, R. (2001). A cognitive-behavioral model of pathological Internet use. *Computers in Human Behavior*, 17(2), 187-195. doi:10.1016/s0747-5632(00)00041-8
- Davis, M. A., Anthony, D. L., & Pauls, S. D. (2015). Seeking and receiving social support on Facebook for surgery. *Social Science & Medicine*, 131, 40-47. doi:10.1016/j.socscimed.2015.02.038
- De Choudhury, M., & De, S. (2014). Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *ICWSM*. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8075/8107>
- De Choudhury, M., & Kiciman, E. (2017). The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. *International AAAI Conference on Weblogs and Social Media* (Vol. 2017, p. 32). NIH Public Access. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5565730/pdf/nihms891351.pdf>
- De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. doi:10.1145/2858036.2858207
- De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., & Nielsen, R. C. (2017). Gender and cross-cultural differences in social media disclosures of mental illness. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. doi:10.1145/2998181.2998220
- DeAndrea, D. C., & Anthony, J. C. (2013). Online peer support for mental health problems in the United States: 2004–2010. *Psychological Medicine*, 43(11), 2277-2288. doi:10.1017/s0033291713000172
- Dehghani, M., Johnson, K., Garten, J., Balasubramanian, V., Singh, A., & Shankar, Y. et al. (2017). TACIT: An Open-Source Text Analysis, Crawling and Interpretation Tool. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2660651
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78. doi:10.1145/2347736.2347755
- Dutta, S., Ma, J., & De Choudhury, M. (2018). Measuring the impact of anxiety on online social interactions. In *Twelfth International AAAI Conference on Web and Social Media*. Retrieved from http://www.munmund.net/pubs/Anxiety_SocialInt_ICWSM18.pdf
- Eskreis-Winkler, L., Fishbach, A., & Duckworth, A. L. (2018). Dear Abby: Should I give advice or receive it? *Psychological Science*, 29(11), 1797-1806. doi:10.1177/0956797618795472

- Feng, B., & Burleson, B. R. (2008). The effects of argument explicitness on responses to advice in supportive interactions. *Communication Research*, 35(6), 849-874. doi:10.1177/0093650208324274
- Feng, B., Jang, J. Y., Kim, I., & Wang, B. (2017). Explaining the effectiveness of the integrated model of advice-giving in supportive interactions: The mediating roles of politeness and normativeness. *Communication Quarterly*, 65(4), 402-418. doi:10.1080/01463373.2017.1279676
- Feng, B., Li, S., & Li, N. (2016). Is a profile worth a thousand words? How online support-seeker's profile features may influence the quality of received support messages. *Communication Research*, 43(2), 253-276. doi:10.1177/0093650213510942
- Feng, B., & MacGeorge, E. L. (2006). Predicting receptiveness to advice: Characteristics of the problem, the advice-giver, and the recipient. *Southern Communication Journal*, 71(1), 67-85. doi:10.1080/10417940500503548
- Feng, B., & MacGeorge, E. L. (2010). The influences of message and source factors on advice outcomes. *Communication Research*, 37(4), 553-575. doi:10.1177/0093650210368258
- Filieri, R. (2016). What makes an online consumer review trustworthy? *Annals of Tourism Research*, 58, 46-64. doi:10.1016/j.annals.2015.12.019
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221-233. doi:10.1037/h0057532
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5), 75-174. doi:10.1016/j.physrep.2009.11.002
- Fraga, B.S., Couto da Silva, A. P., & Murai, F. (2018). Online social networks in health care: A study of mental disorders on Reddit. 2018 *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. doi:10.1109/wi.2018.00-36
- Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (2018). "Let Me Tell You About Your Mental Health!" Contextualized Classification of Reddit Posts to DSM-5 for Web-Based Intervention. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18*, 1-10. doi:10.1145/3269206.3271732
- Gino, F., Brooks, A. W., & Schweitzer, M. E. (2012). Anxiety, advice, and the ability to discern: Feeling anxious motivates individuals to seek and use advice. *Journal of Personality and Social Psychology*, 102(3), 497-512. doi:10.1037/a0026413
- Giota, K. G., & Kleftras, G. (2014). The discriminant value of personality, motivation, and online relationship quality in predicting attraction to online social support on Facebook. *International Journal of Human-Computer Interaction*, 30(12), 985-994. doi:10.1080/10447318.2014.925770
- Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T., Dobson, R., & Dutta, R. (2017). Characterisation of mental health conditions in

- social media using Informed Deep Learning. *Scientific Reports*, 7(1), 45141. doi:10.1038/srep45141
- Goldsmith, D. J. (1999). Content-based resources for giving face sensitive advice in troubles talk episodes. *Research on Language and Social Interaction*, 32(4), 303-336. doi:10.1207/s15327973rls3204_1
- Goldsmith, D. J. (2004). *Communicating social support*. Cambridge, NY: Cambridge University Press.
- Goldsmith, D. J., & Fitch, K. (1997). The normative context of advice as social support. *Human Communication Research*, 23(4), 454-476. doi:10.1111/j.1468-2958.1997.tb00406.x
- Goldsmith, D. J., & MacGeorge, E. (2000). The impact of politeness and relationship on perceived quality of advice about a problem. *Human Communication Research*, 26(2), 234-263. doi:10.1111/j.1468-2958.2000.tb00757.x
- Goldsmith, D. J., McDermott, V. M., & Alexander, S. C. (2000). Helpful, supportive and sensitive: Measuring the evaluation of enacted social support in personal relationships. *Journal of Social and Personal Relationships*, 17(3), 369-391. doi:10.1177/0265407500173004
- Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2009). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1), 3-19. doi:10.1177/0093650209351468
- González-Bailón, S. (2013). Social science in the era of big data. *Policy & Internet*, 5(2), 147-160. doi:10.1002/1944-2866.poi.328
- Goodman, M. W. (2019). A Python Library for Deep Linguistic Resources. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)* (pp. 1-7). IEEE.
- Graham, S. M., Huang, J. Y., Clark, M. S., & Helgeson, V. S. (2008). The positives of negative emotions: Willingness to express negative emotions promotes relationships. *Personality and Social Psychology Bulletin*, 34(3), 394-406. doi:10.1177/0146167207311281
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1), 5228-5235. doi:10.1073/pnas.0307752101
- Gross, E. F., Juvonen, J., & Gable, S. L. (2002). Internet use and well-being in adolescence. *Journal of Social Issues*, 58(1), 75-90. doi:10.1111/1540-4560.00249
- Harper, V. B., & Harper, E. J. (2006). Understanding student self-disclosure typology through blogging. *The Qualitative Report*, 11(2), 251-261. Retrieved from <https://nsuworks.nova.edu/tqr/vol11/iss2/3>
- Harrison, S., & Barlow, J. (2009). Politeness strategies and advice-giving in an online arthritis workshop. *Journal of Politeness Research. Language, Behaviour, Culture*, 5(1). doi:10.1515/jplr.2009.006
- Hinson, J. A., & Swanson, J. L. (1993). Willingness to seek help as a function of self-disclosure and problem severity. *Journal of Counseling & Development*, 71(4), 465-470. doi:10.1002/j.1556-6676.1993.tb02666.x

- Hofmann, D. A., Lei, Z., & Grant, A. M. (2009). Seeking help in the shadow of doubt: The sensemaking processes underlying how nurses decide whom to ask for advice. *Journal of Applied Psychology*, 94(5), 1261-1274. doi:10.1037/a0016557
- House, J. (1981). *Work stress and social support*. Reading, Mass.: Addison-Wesley Pub. Co
- Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2), 561-569. doi:10.1016/j.chb.2011.11.001
- Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*. Retrieved from <http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>
- Hwang, K. O., Ottenbacher, A. J., Lucke, J. F., Etchegaray, J. M., Graham, A. L., Thomas, E. J., & Bernstam, E. V. (2011). Measuring social support for weight loss in an Internet weight loss community. *Journal of Health Communication*, 16(2), 198-211. doi:10.1080/10810730.2010.535106
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London and New York: Continuum International Publishing Group Ltd.
- Ji, S., Yu, C. P., Fung, S., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018, 1-10. doi:10.1155/2018/6157249
- Joyce, E., & Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3), 723-747. doi:10.1111/j.1083-6101.2006.00033.x
- Kane, A. A., & Rink, F. (2015). How newcomers influence group utilization of their knowledge: Integrating versus differentiating strategies. *Group Dynamics: Theory, Research, and Practice*, 19(2), 91-105. doi:10.1037/gdn0000024
- Karrer, B., & Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1). doi:10.1103/physreve.83.016107
- Kasunic, A., & Kaufman, G. (2018). "At Least the Pizzas You Make Are Hot": Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe. In *ICWSM* (pp. 161-170). Retrieved from <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17870/17007>
- Kern, M., Park, G., Eichstaedt, J., Schwartz, H., Sap, M., Smith, L., & Ungar, L. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, 21(4), 507-525. doi:10.1037/met0000091

- Klein, C., Clutton, P., & Polito, V. (2018). Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in Psychology*, 9. doi:10.3389/fpsyg.2018.00189
- Kornfield, R., Sarma, P. K., Shah, D. V., McTavish, F., Landucci, G., Peromashko, K., & Gustafson, D. H. (2018). Detecting recovery problems just in time: Application of automated linguistic analysis and supervised machine learning to an online substance abuse forum. *Journal of Medical Internet Research*, 20(6), e10136. doi:10.2196/10136
- Kim, J., Weinstein, E. C., & Selman, R. L. (2016). Romantic relationship advice from anonymous online helpers: The peer support adolescents exchange. *Youth & Society*, 49(3), 369-392. doi:10.1177/0044118x15604849
- Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, 21(4), 493-506. doi:10.1037/met0000105
- Kumar, M., Dredze, M., Coppersmith, G., & De Choudhury, M. (2015). Detecting changes in suicide content manifested in social media following celebrity suicides. *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 85-94. doi:10.1145/2700171.2791026
- Lagoe, C., & Atkin, D. (2015). Health anxiety in the digital age: An exploration of psychological determinants of online health information seeking. *Computers in Human Behavior*, 52, 484-491. doi:10.1016/j.chb.2015.06.003
- Lambiotte, R., & Kosinski, M. (2014). Tracking the Digital Footprints of Personality. *Proceedings Of The IEEE*, 102(12), 1934-1939. doi:10.1109/jproc.2014.2359054
- Landers, R., Brusso, R., Cavanaugh, K., & Collmus, A. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, 21(4), 475-492. doi:10.1037/met0000081
- Lawlor, A., & Kirakowski, J. (2014). Online support groups for mental health: A space for challenging self-stigma or a means of social avoidance? *Computers in Human Behavior*, 32, 152-161. doi:10.1016/j.chb.2013.11.015
- Ledbetter, A. M. (2007). Chronemic cues and sex differences in relational e-mail: Perceiving immediacy and supportive message quality. *Social Science Computer Review*, 26(4), 466-482. doi:10.1177/0894439308314812
- Lee, B. W., & Stapinski, L. A. (2012). Seeking safety on the internet: Relationship between social anxiety and problematic internet use.

- Journal of Anxiety Disorders*, 26(1), 197-205.
doi:10.1016/j.janxdis.2011.11.001
- Leimeister, J.M., Schweizer, K., Leimeister, S., & Krcmar, H. (2008). Do virtual communities matter for the social support of patients?. *Information Technology & People*, 21(4), 350-374.
doi:10.1108/09593840810919671
- Leung, L. (2007). Stressful Life Events, Motives for Internet Use, and Social Support Among Digital Kids. *CyberPsychology & Behavior*, 10(2), 204-214. doi:10.1089/cpb.2006.9967
- Livingstone, S., & Helsper, E. (2007). Gradations in digital inclusion: Children, young people and the digital divide. *New Media & Society*, 9(4), 671-696. doi:10.1177/1461444807080335
- Lo, J. (2019). Exploring the buffer effect of receiving social support on lonely and emotionally unstable social networking users. *Computers in Human Behavior*, 90, 103-116. doi:10.1016/j.chb.2018.08.052
- Locher, M. A. (2006). Polite behavior within relational work: The discursive approach to politeness. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 25(3), 249-267.
doi:10.1515/multi.2006.015
- Locher, M. A., & Hoffmann, S. (2006). The emergence of the identity of a fictional expert advice-giver in an American Internet advice column. *Text & Talk - An Interdisciplinary Journal of Language, Discourse Communication Studies*, 26(1), 69-106. doi:10.1515/text.2006.004
- Li, S., Coduto, K. D., & Morr, L. (2019). Communicating social support online: The roles of emotional disclosures and gender cues in support provision. *Telematics and Informatics*, 39, 92-100. doi:10.1016/j.tele.2019.02.004
- Li, Y., Mihalcea, R., & Wilson, S. R. (2018). Text-Based Detection and Understanding of Changes in Mental Health. *Lecture Notes in Computer Science*, 176-188. doi:10.1007/978-3-030-01159-8_17
- Lu, Y., Mei, Q., & Zhai, C. (2010). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 14(2), 178-203. doi:10.1007/s10791-010-9141-9
- Lyons, M., Aksayli, N. D., & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 87, 207-211. doi:10.1016/j.chb.2018.05.035
- MacGeorge, E. L., Feng, B., & Guntzviller, L. M. (2016a). Advice: expanding the communication paradigm. *Annals of the International Communication Association*, 40(1), 213-243.
doi:10.1080/23808985.2015.11735261
- MacGeorge, E. L., Guntzviller, L. M., Brisini, K. S., Bailey, L. C., Salmon, S. K., Severen, K., ... Cummings, R. D. (2016b). The influence of emotional support quality on advice evaluation and outcomes. *Communication Quarterly*, 65(1), 80-96.
doi:10.1080/01463373.2016.1176945

- MacGeorge, E. L., Lichtman, R. M., & Pressey, L. C. (2002). The evaluation of advice in supportive interactions: Facework and contextual factors. *Human Communication Research*, 28(3), 451-463. doi:10.1111/j.1468-2958.2002.tb00815.x
- Malloch, Y. Z., & Taylor, L. D. (2018). Emotional self-disclosure in online breast cancer support groups: Examining theme, reciprocity, and linguistic style matching. *Health Communication*, 34(7), 764-773. doi:10.1080/10410236.2018.1434737
- Mansell, W., & Lam, D. (2006). "I Won't Do What You Tell Me!": Elevated mood and the assessment of advice-taking in euthymic bipolar I disorder. *Behaviour Research and Therapy*, 44(12), 1787-1801. doi:10.1016/j.brat.2006.01.002
- Massanari, A. (2016). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3), 329-346. doi:10.1177/1461444815608807
- Mattsson, S., Olsson, E. M., Alfonsson, S., Johansson, B., & Carlsson, M. (2015). Measuring use of health-related support on the internet: Development of the Health Online Support Questionnaire (HOSQ). *Journal of Medical Internet Research*, 17(11), e266. doi:10.2196/jmir.4425
- Mazzoni, E., Baiocco, L., Cannata, D., & Dimas, I. (2016). Is internet the cherry on top or a crutch? Offline social support as moderator of the outcomes of online social support on problematic internet use. *Computers in Human Behavior*, 56, 369-374. doi:10.1016/j.chb.2015.11.032
- McDonald, D., & Woodward-Kron, R. (2016). Member roles and identities in online support groups: Perspectives from corpus and systemic functional linguistics. *Discourse & Communication*, 10(2), 157-175. doi:10.1177/1750481315615985
- McKenna, K. Y., & Bargh, J. A. (1999). Causes and consequences of social interaction on the Internet: A conceptual framework. *Media Psychology*, 1(3), 249-269. doi:10.1207/s1532785xmep0103_4
- Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, 51(7), 656-667. doi:10.1002/eat.22878
- Morrison, M., & Roese, N. J. (2011). Regrets of the typical American: Findings from a nationally representative sample. *Social Psychological and Personality Science*, 2(6), 576-583. doi:10.1177/1948550611401756
- Morrow, P. R. (2006). Telling about problems and giving advice in an Internet discussion forum: some discourse features. *Discourse Studies*, 8(4), 531-548. doi:10.1177/1461445606061876
- Mulshine, M. (2014). *This Insane Relationship Advice Subreddit Is Basically an Online Soap Opera*. *Observer*. Retrieved 22 April 2018, from

<http://observer.com/2014/07/this-insane-relationship-advice-subreddit-is-basically-an-online-soap-opera/>

- Newman, M. E. (2004). Analysis of weighted networks. *Physical Review E*, 70(5). doi:10.1103/physreve.70.056131
- Niemz, K., Griffiths, M., & Banyard, P. (2005). Prevalence of pathological Internet use among university students and correlations with self-esteem, the General Health Questionnaire (GHQ), and disinhibition. *CyberPsychology & Behavior*, 8(6), 562-570. doi:10.1089/cpb.2005.8.562
- Olson, R., & Neal, Z. (2015). Navigating the massive world of reddit: using backbone networks to map user interests in social media. *PeerJ Computer Science*, 1(e4), 1-13. doi:10.7717/peerj-cs4
- Pappa, G. L., Cunha, T. O., Bicalho, P. V., Ribeiro, A., Couto Silva, A. P., Meira Jr, W., & Beleigoli, A. M. (2017). Factors associated with weight change in online weight management communities: A case study in the loseit reddit community. *Journal of Medical Internet Research*, 19(1), e17. doi:10.2196/jmir.581
- Park, A., & Conway, M. (2018). Harnessing Reddit to understand the written-communication challenges experienced by individuals with mental health disorders: Analysis of texts from mental health communities. *Journal of Medical Internet Research*, 20(4), e121. doi:10.2196/jmir.8219
- Park, M., Cha, C., & Cha, M. (2012, August). Depressive moods of users portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)* (Vol. 2012, pp. 1-8). New York, NY: ACM. Retrieved from <https://pdfs.semanticscholar.org/8dd5/8913bd343f4ef23b8437b24e152d3270cdaf.pdf>
- Pavalanathan, U., & De Choudhury, M. (2015). Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 315-321). ACM. doi:10.1145/2740908.2743049
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?Sequence=3
- Pennebaker, J. W., Chung, C. K., Frazee, J., Laverne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, 9(12), e115844. doi:10.1371/journal.pone.0115844

- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1296-1312. doi: 10.1037/0022-3514.77.6.1296
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 547-577. doi:10.1146/annurev.psych.54.101601.145041
- Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2), 291-301. doi:10.1037/0022-3514.85.2.291
- Poston, R. S., Akbulut, A. Y., & Looney, C. A. (2009). Examining the effects of self-efficacy, computerised sources and perceived source credibility in decisions to take online advice. *International Journal of Electronic Business*, 7(4), 321. doi:10.1504/ijeb.2009.027264
- r/ListofSubreddits wiki*. (2017). *Reddit.com*. Retrieved 11 November 2018, from <https://www.reddit.com/r/ListOfSubreddits/wiki/defaults>
- r/relationships wiki*. (2019). *Reddit.com*. Retrieved 12 June 2019, from <https://www.reddit.com/r/relationships/wiki/index>
- Rains, S. A. (2014). The implications of stigma and anonymity for self-disclosure in health blogs. *Health Communication*, 29(1), 23-31. doi:10.1080/10410236.2012.714861
- Rains, S. A., & Wright, K. B. (2016). Social Support and Computer-Mediated Communication: A State-Of-The-Art Review and Agenda for Future Research. *Annals of the International Communication Association*, 40(1), 175-211. doi:10.1080/23808985.2015.11735260
- Rains, S. A., Tsetsi, E., Akers, C., Pavlich, C. A., & Appelbaum, M. (2018). Factors Influencing the Quality of Social Support Messages Produced Online: The Role of Responsibility for Distress and Others' Support Attempts. *Communication Research*, 1-21. doi: 10.1177/009365021879637
- Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133-142). Retrieved from <https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf>
- Record, R. A., Silberman, W. R., Santiago, J. E., & Ham, T. (2018). I sought it, I Reddit: Examining health information engagement behaviors among Reddit users. *Journal of Health Communication*, 23(5), 470-476. doi:10.1080/10810730.2018.1465493
- redditlist.com*. (2019). *redditlist*. Retrieved 12 June 2019, from <http://redditlist.com/all>
- Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. doi: 0.13140/2.1.2393.1847

- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15.*, 399-408. doi: 10.1145/2684822.2685324
- Ruble, R. A. (2011). The communication of advice on an online message board for language assistants in France. *Journal of Language and Social Psychology*, 30(4), 396-420. doi:10.1177/0261927x11416204
- Schaub, M. T., Delvenne, J., Rosvall, M., & Lambiotte, R. (2017). The many facets of community detection in complex networks. *Applied Network Science*, 2(1). doi:10.1007/s41109-017-0023-6
- Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The Annals of the American Academy of Political and Social Science*, 659(1), 78-94. doi:10.1177/0002716215569197
- Serrano, M. A., Boguna, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16), 6483-6488. doi:10.1073/pnas.0808904106
- Sharma, E., & De Choudhury, M. (2018). Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. doi:10.1145/3173574.3174215
- Shaw, L. H., & Gant, L. M. (2002). In defense of the Internet: The relationship between Internet communication and depression, loneliness, self-esteem, and perceived social support. *CyberPsychology & Behavior*, 5(2), 157-171. doi:10.1089/109493102753770552
- Shelton, M., Lo, K., & Nardi, B. (2015). Online media forums as separate social lives: A qualitative study of disclosure within and beyond Reddit. *iConference 2015 Proceedings*. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/73676/75_ready.pdf?sequence=2&isAllowed=y
- Shen, J. H., & Rudzicz, F. (2017). Detecting anxiety through Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 58-65. doi:10.18653/v1/w17-3107
- Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63-70. Retrieved from <http://www.aclweb.org/anthology/W14-3110>
- Sillence, E. (2010). Seeking out very like-minded others: Exploring trust and advice issues in an online health support group. *International Journal of Web Based Communities*, 6(4), 376. doi:10.1504/ijwbc.2010.035840
- Smith, S. (2018). *What Our Obsession With Reddit Relationships Reveal About Us. Bitch Media*. Retrieved 22 April 2018, from

<https://www.bitchmedia.org/article/the-weird-and-wild-world-of-reddit-relationships>

- Smithson, J., Sharkey, S., Hewis, E., Jones, R. B., Emmens, T., Ford, T., & Owens, C. (2011). Membership and boundary maintenance on an online self-harm forum. *Qualitative Health Research*, 21(11), 1567-1575. doi:10.1177/1049732311413784
- Soldaini, L., Walsh, T., Cohan, A., Han, J., & Goharian, N. (2018). Helping or hurting? Predicting changes in users' risk of self-harm through online community interactions. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. doi:10.18653/v1/w18-0621
- Spottswood, E. L., Walther, J. B., Holmstrom, A. J., & Ellison, N. B. (2013). Person-centered emotional support and gender attributions in computer-mediated communication. *Human Communication Research*, 39(3), 295-316. doi:10.1111/hcre.12006
- Stanton, A. M., Boyd, R. L., Pulverman, C. S., & Meston, C. M. (2015). Determining women's sexual self-schemas through advanced computerized text analysis. *Child Abuse & Neglect*, 46, 78-88. doi:10.1016/j.chiabu.2015.06.003
- Stirman, S.W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4), 517-522. doi:10.1097/00006842-200107000-00001
- Suzuki, L. K., & Calzo, J. P. (2004). The search for peer advice in cyberspace: An examination of online teen bulletin boards about health and sexuality. *Journal of Applied Developmental Psychology*, 25(6), 685-698. doi:10.1016/j.appdev.2004.09.002
- Tamersoy, A., De Choudhury, M., & Chau, D. H. (2015). Characterizing smoking and drinking abstinence from social media. *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*. doi:10.1145/2700171.2791247
- Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016, April). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web* (pp. 613-624). International World Wide Web Conferences Steering Committee. doi: 10.1145/2872427.2883081
- Tardy, C. (1985). Social support measurement. *American Journal Of Community Psychology*, 13(2), 187-202. doi:10.1007/bf00905728
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1), 24-54. doi:10.1177/0261927x09351676
- Testing a new sign up experience: changelog*. (2017). reddit. Retrieved 11 November 2018, from

- https://www.reddit.com/r/changelog/comments/65x4a7/testing_a_new_sign_up_experience/
- Thayer, S. E., & Ray, S. (2006). Online communication preferences across age, gender, and duration of Internet use. *CyberPsychology & Behavior*, 9(4), 432-440. doi:10.1089/cpb.2006.9.432
- Thomas, S. W. (2011). Mining software repositories using topic models. In *Proceedings of the 33rd International Conference on Software Engineering* (pp. 1138-1139). ACM. Retrieved from http://cs.queensu.ca/~sthomas/data/Thomas_2012_TR2012-586.pdf
- Tichon, J., & Shapiro, M. (2003). The Process of Sharing Social Support in Cyberspace. *Cyberpsychology & Behavior*, 6(2), 161-170. doi:10.1089/109493103321640356
- Toma, C. L., & D'Angelo, J. D. (2014). Tell-tale words: Linguistic cues used to infer the expertise of online medical advice. *Journal of Language and Social Psychology*, 34(1), 25-45. doi:10.1177/0261927x14554484
- Tomprou, M., Dabbish, L., Kraut, R. E., & Liu, F. (2019). Career mentoring in online communities: Seeking and receiving advice from an online community. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. doi:10.1145/3290605.3300883
- Tost, L. P., Gino, F., & Larrick, R. P. (2012). Power, competitiveness, and advice taking: Why the powerful don't listen. *Organizational Behavior and Human Decision Processes*, 117(1), 53-65. doi:10.1016/j.obhdp.2011.10.001
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*. doi:10.3115/1073445.1073478
- Tufekci, Z. (2014, May). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*. Retrieved from <https://arxiv.org/ftp/arxiv/papers/1403/1403.7400.pdf>
- Update To Search API: changelog. (2018). *reddit*. Retrieved 11 November 2018, from https://www.reddit.com/r/changelog/comments/7tus5f/update_to_search_api/
- Van Mieghem, P. (2011). Human psychology of common appraisal: The reddit score. *IEEE Transactions on Multimedia*, 13(6), 1404-1406. doi:10.1109/tmm.2011.2165054
- Van Swol, L. M. (2011). Forecasting another's enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, 27(1), 103-120. doi:10.1016/j.ijforecast.2010.03.002

- Van Swol, L. M., MacGeorge, E. L., & Prah, A. (2017). Advise with permission?: The effects of advice solicitation on advice outcomes. *Communication Studies*, 68(4), 476-492. doi:10.1080/10510974.2017.1363795
- Vinh, N. X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. doi:10.1145/1553374.1553511
- Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in neural information processing systems* (pp. 1973-1981). Retrieved from <http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf>
- Walther, J. B., & Boyd, S. (2002). Attraction to computer-mediated social support. *Communication technology and society: Audience adoption and uses*, 153-188. Retrieved from <https://msu.edu/~jwalther/docs/support.html>
- Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior*, 23(5), 2538-2557. doi:10.1016/j.chb.2006.05.002
- Wang, Y., Burke, M., & Kraut, R. E. (2013). Gender, topic, and audience response: An analysis of user-generated content on facebook. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. doi:10.1145/2470654.2470659
- Wang, Z., & Jurgens, D. (2018). It's going to be okay: Measuring Access to Support in Online Communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 33-45). Retrieved from <https://www.aclweb.org/anthology/D18-1004>
- Wang, Y., Kraut, R., & Levine, J. M. (2012). To stay or leave?: The relationship of emotional and informational support to commitment in online health support groups. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. doi:10.1145/2145204.2145329
- Wang, Y., Kraut, R. E., & Levine, J. M. (2015). Eliciting and Receiving Online Support: Using Computer-Aided Content Analysis to Examine the Dynamics of Online Social Support. *Journal of Medical Internet Research*, 17(4), 1-23. doi:10.2196/jmir.3558
- Weinstein, E. C., & Selman, R. L. (2014). Digital stress: Adolescents' personal accounts. *New Media & Society*, 18(3), 391-409. doi:10.1177/1461444814543989
- Wheless, L. R., & Grotz, J. (1977). The measurement of trust and its relationship to self-disclosure. *Human Communication Research*, 3(3), 250-257. doi:10.1111/j.1468-2958.1977.tb00523.x

- White, M., & Dorman, S. M. (2001). Receiving social support online: Implications for health education. *Health Education Research*, 16(6), 693-707. doi:10.1093/her/16.6.693
- Wierzbicka, A. (1985). Different cultures, different languages, different speech acts. *Journal of Pragmatics*, 9(2-3), 145-178. doi:10.1016/0378-2166(85)90023-2
- Wolohan, J. T., Hiraga, M., Mukherjee, A., Sayyed, Z. A., & Millard, M. (2018, August). Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models* (pp. 11-21). Retrieved from <https://www.aclweb.org/anthology/W18-4102>
- Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1), 1-13. doi:10.1016/j.obhdp.2003.08.002
- Yaniv, I., & Choshen-Hillel, S. (2012). When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advice-taking. *Journal of Experimental Social Psychology*, 48(5), 1022-1028. doi:10.1016/j.jesp.2012.03.016
- Zhao, K., Stylianou, A. C., & Zheng, Y. (2013). Predicting users' continuance intention in virtual communities: The dual intention-formation processes. *Decision Support Systems*, 55(4), 903-910. doi:10.1016/j.dss.2012.12.026
- Zhong, W., An, H., Gao, X., & Sun, X. (2014). The evolution of communities in the international oil trade network. *Physica A: Statistical Mechanics and its Applications*, 413, 42-52. doi:10.1016/j.physa.2014.06.055

APPENDIX A: TOPIC MODELLING ON POSTS AND COMMENT GROUPS.

The 70520 posts and 68925 comment groups were cleaned into a form usable for topic modelling (i.e., continuous prose into a simple list of words). The raw text was processed as follows: (1) newline and carriage return characters (`\n`, `\r`) were removed, (2) Reddit usernames and URLs were removed, (3) punctuation and numbers were removed, (4) some short-forms converted to Standard English (e.g. ‘bf’ to ‘boyfriend’), (5) names of people¹¹ and common words (i.e., stop-words) were removed, and (6) words that were just one character were dropped. For standardisation, all words were lowercased. Bigrams (i.e., two consecutive words – for example, “red flags”) that appeared at least 20 times across all 70520 posts were appended to their respective posts. The same was done for the comment groups.

A dictionary of words that appeared in the processed 70520 posts was then created by collating all available words and then filtering out words on the extreme ends of the frequency count – those that appeared in less than 20 posts or appeared in more than 40% of posts. This filtering was done to improve the resulting quality of the topic model (Thomas, 2011). The same was done for the comment groups.

LDA was set to run from $k=2$ to $k=60$ topics for the post and comment data separately, in steps of 2 (i.e., $k=2, 4, 6 \dots 60$) with the α parameter set to ‘auto’. The α parameter is the prior for the distributions of topics in documents and affects the sparsity of the solution; a smaller α value will result in fewer dominant (high-probability) topics within each document (Lu, Mei & Zhai, 2011; Kosinski et al., 2016). Typically, LDA is run with a symmetric α prior set by the researcher (Wallach, Mimno & McCallum, 2009); setting it to $50/k$ as

¹¹ A list of person names was created by running the Stanford Named Entity Recognizer (Finkel, Grenager & Manning, 2005) across all 70520 posts and 68925 comment groups. Entities identified as people were saved into a list which was manually inspected. Words that were misidentified as names of people (e.g. Runescape, Gonewild, aspie) were removed from the list.

did Griffiths and Steyvers (2004) is a common recommendation. Setting the α parameter to ‘auto’ in this paper allowed an asymmetric prior to be learned from the given data instead. Wallach et al., (2009) found that such a setting resulted in better learned topic quality and also had the beneficial effect of clustering stop-words specific to the document collection into a few common topics. Other topics would hence be freer of such document-collection specific words and more clearly expressive of unique themes (Wallach et al., 2009).

A good topic model would be one which has human-interpretable topics. However, selecting the optimal k in LDA is a difficult task with no homogenised protocol available (Wang et al., 2013). The gold standard – using human judges – is expensive and laborious to obtain (Roder, Both & Hinneburg, 2015). Following Gaur et al., (2018), the current paper thus turned to using topic coherence, an automated quantitative measure of topic understandability (Roder et al., 2015), to assist in the selection of k . The coherence measure used in the current study was C_v , which was proposed and shown by Roder et al., (2015) to correlate rather well with ratings collected from human judges. At each k , C_v was computed. The C_v scores for all the topic models were then plotted and the higher-scoring models selected for manual inspection. The final selected k -topic model was thus one that scored highly on the automated C_v measure and which was checked to be human-interpretable. Tables A1 and A2 show the retrieved topic and their relevant keywords.

Table A1. 28 topics retrieved from 70520 posts and corresponding keywords.

Topic Label	Top 10 words ranked by relevance metric
General forum language	advice, however, situation, issue, part, recently, issues, past, due, well
Hedging/quantifiers	kind, pretty, asked, little, thing, bit, sure, ask, well, say
Past partners	started, ago, wanted, felt, weeks, ex, broke, ended, month, made
Communication	boyfriend, always, says, wants, talk, sometimes, every, makes, tell, try

Romantic feelings	love, life, someone, person, anyone, feeling, else, relationships, happy, feelings
Fighting/arguments	upset, angry, fight, mad, saying, wrong, argument, behavior, explain, comments
Mental health	anxiety, depression, help, abusive, mental, depressed, health, therapy, support, therapist
Living arrangements (hygiene/chores)	home, dog, take, house, clean, day, driving, drive, food, watch
Texting/calling	day, text, night, week, next, texted, texting, texts, busy, days
Moving	move, live, moved, living, moving, city, home, away, place, stay
Dating	girl, guy, dating, date, guys, girls, met, interested, dates, women
Social media & trust issues	trust, phone, facebook, cheating, cheated, found, messages, pictures, lying, media
Family	mom, dad, family, parents, mother, sister, brother, father, mum, cousin
Finances	money, pay, rent, paying, paid, afford, car, financial, buy, financially
Socializing	people, group, uncomfortable, social, everyone, friendly, event, coworker, awkward, co
Friends	friend, best, best_friend, friendship, mutual, close, female, mutual_friends, gay, mutual_friend
School	school, college, high, high_school, class, classes, semester, summer, graduated, program
Marital	husband, kids, wife, baby, son, daughter, child, children, divorce, married
Living arrangements (space)	room, bed, sleep, door, car, cat, bathroom, bedroom, house, night
Sexual	sex, sexual, sexually, porn, fwb, intimacy, kiss, attracted, intimate, kissing
Long-distance relationships	girlfriend, distance, long_distance, term, long, long_term, ldr, current_girlfriend, break, internship

Jobs	job, company, boss, office, working, quit, business, team, career, jobs
Trips/occasions	weekend, trip, birthday, vacation, visit, holiday, plans, hotel, spend, planned
Drinking/drugs/religion	drunk, drinking, drink, bar, drinks, esteem, alcohol, self, church, self_esteem
Hobbies/interests	play, games, playing, game, video, gift, gifts, plays, gaming, music
Body image	weight, eating, gym, body, wear, dress, fat, eat, naked, gained
Miscellaneous ¹²	university, country, study, studying, language, uni, english, abroad, project, culture
Getting married	wedding, married, marriage, engaged, getting_married, marry, ring, spring_break, breaker, deal_breaker

Table A2. 36 topics retrieved from 68925 comment groups and corresponding keywords.

Topic Label	Top 10 words ranked by relevance metric
General forum language	good, life, love, feel, people, much, things, lot, person, still
Communication	say, something, talk, ask, tell, maybe, might, probably, see, could
Cutting contact/poor behaviour	leave, stop, shit, back, someone, sorry, deserve, respect, cut, guy
Romantic: long-term/long distance	together, break, long, months, end, two, move, future, ready, wants
Interpretation of issues	sounds, boyfriend, sounds_like, problem, issue, sound, partner, needs, sound_like, upset
Taking action	take, call, use, keep, boss, car, make_sure, drive, check, order
Mental health: therapy	help, therapy, support, therapist, needs, fix, care, issues, seek, professional

¹² Though this topic appeared interpretable at first (e.g. university studies), examining the posts which had higher proportions of it showed that it was actually rather heterogenous.

Family	parents, mom, mother, dad, sister, brother, father, daughter, family, son
Marriage	husband, wife, marriage, kids, married, divorce, children, marry, lawyer, divorced
Activity scheduling	week, spend, day, night, hours, spend_time, days, home, busy, weekend
Physical attractiveness	women, self, attractive, esteem, men, self_esteem, insecure, joke, body, confidence
Dating	date, dating, guy, girl, interested, meet, someone, dates, guy, met
Finances	money, pay, paying, financial, buy, afford, bills, finances, paid, account
Food/eating preparation	food, weight, eat, eating, chores, clean, cleaning, cook, cooking, diet
Careers/school	job, school, college, career, work, jobs, high_school, student, hr, university
Mental health: disorder/diagnosis/professional	health, mental, anxiety, depression, doctor, mental_health, medical, illness, medication, disorder
Friends: current issues	friend, friendship, ex, friends, best_friend, crush, feelings, best, exes, platonic
Hobbies/interests	hobbies, hobby, interests, new, fun, activities, enjoy, sports, class, practice
Age gap	age, old, red, year_old, year, flag, older, red_flag, flags, ring
Lying/secrets	lying, lie, lied, tell, truth, business, lies, coworker, trust, secret
Sleeping arrangements	room, sleep, bed, house, cat, roommates, sleeping, cats, home, bedroom
Sex	sex, sexual, porn, sexually, intimacy, monogamous, libido, sex_drive, attracted, intimate

Friends: making	friends, people, group, social, hang, friendships, circle, groups, friend_group, meet
Family: moving/extended/loss	family, trip, country, vacation, grief, aunt, died, cousin, travel, visit
Cheating	girlfriend, cheating, trust, cheated, cheat, cheater, girl, tested, dump, poly
Special occasions	wedding, birthday, party, fiancé, invite, fianc, invited, christmas, holiday, celebrate
Electronic communication	phone, block, text, media, contact, send, social_media, facebook, message, messages
Social media/app usage	online, app, snooping, internet, search, dance, delete, music, photo, dancing
Abuse	abusive, abuse, police, abused, violence, abuser, violent, abusing, report, cps
Love compatibility	gift, gifts, letter, religious, church, religion, affection, beliefs, love, flowers
Space/touch	space, cry, give_space, hands, touch, hug, touching, crying, ice, cuddle
Lease	living, name, apartment, roommate, live, move, lease, moving, rent, place
Pregnancy/birth control	baby, child, pregnant, kid, birth, pregnancy, birth_control, condoms, babies, test
Gaming	play, game, playing, games, video, language, gaming, watching, watch, tv
Drinking/drugs	drinking, drunk, drink, alcohol, drugs, drug, alcoholic, sober, addict, beer
Miscellaneous	dog, dogs, gay, hair, weed, smoking, smoke, pet, bi, trans

APPENDIX B: CREATION OF THEMATICALLY-SIMILAR GROUPS OF SUBREDDITS.

The initial clustering of the posting histories to form groups of subreddits with similar topical interests closely followed Olson and Neal's (2015) process. Before proceeding with graph creation and community detection, two main considerations were addressed.

Firstly, the activity count threshold to use to identify if a user was interested in a subreddit. Many Reddit users may post/comment just once to a variety of subreddits in their tenure. This once-off behaviour is not particularly informative of interest, may clutter the graph, and make it harder to identify thematically-similar groups of subreddit based on their common users. Olson and Neal (2015) used a threshold of ten; in the current study, a user was considered to have displayed interest in a subreddit if they had posted and/or commented there at least four times in one year. This figure was decided upon by examining the sample's ($n=60434$ identifiable r/relationships posters) characteristics. Discounting non-active accounts from the preceding year (i.e., no activity at all), the median number of posts made by each remaining r/relationships poster ($n=29862$) to different subreddits was calculated. The median of these user-level medians was then calculated, and it was found that r/relationships posters typically posted and/or commented two times per subreddit in the year preceding their first r/relationships post in the 70520-dataset. The larger number of four was thus decided on as a stronger indicator of user interest in a particular subreddit.

Secondly, prior to mid-2017 ("Testing a new sign up experience : changelog", 2017), Reddit had a default subreddit system which automatically subscribed new users to a site-curated set of subreddits. These subreddits tended to be broad (e.g. r/pics) and may be less indicative of specific user interests. 61 such subreddits were removed based on a user-compiled list of default subreddits (r/ListofSubreddits wiki, 2017). Note that r/relationships

itself was not included in this graph –posts and comments made on it in the year preceding the user’s first r/relationships post in the 70520-dataset were counted separately.

To form the graph of subreddits, the current study began by creating a bipartite graph. User nodes were connected via unweighted links to subreddit nodes they had shown interest in. As mentioned earlier, users were taken to have been interested in a subreddit if they had posted and/or commented there at least four times, and the 61 default subreddits were excluded. The bipartite graph was then simplified by projecting it into an undirected unipartite graph. In this simplified graph, only subreddit nodes were left. The subreddit nodes were connected to each other by links weighted by the number of common interested users they shared. Isolated subreddit nodes (i.e., shared no common user with any other subreddit) were dropped. There were 10953 subreddit nodes and 709303 links. It was likely that many links were spurious and occurred by chance. Following Olson and Neal (2015), a disparity filter (Serrano, Boguna & Vespignani, 2009) was used to keep only statistically significant links ($\alpha = .001$) that form the network’s main backbone. This reduced the network size to 639 nodes and 2646 links.

Following this, the Louvain community detection algorithm (Blondel et al., 2008) was applied. By optimising the modularity function (Newman, 2004), the Louvain algorithm finds groups of nodes which are more densely connected than would be expected by chance in a random network (Fortunato, 2010; Schaub et al., 2017). No input for the number of clusters to find is required. As optimising the modularity function is computationally difficult, the solutions returned are approximate (Blondel et al., 2008). Thus, the Louvain algorithm was run 1000 times on the graph. To evaluate the consistency of the results, normalised mutual information (NMI), as introduced by Danon, Diaz-Guilera, Duch and Arenas (2005) for use in graph partition comparisons (e.g. Zhong, An, Gao & Sun, 2014), and adjusted mutual information (AMI) (Vinh, Epps & Bailey, 2009), were calculated. The closer NMI and AMI are to 1, the more

similar the compared pair of partitions is (Vinh et al., 2009; Karrer & Newman, 2011). AMI has an adjustment that considers random chance; unadjusted measures were noted to have a penchant for increasing with the number of identified clusters (Vinh et al., 2009). Across all 1000 runs, the results were rather similar – NMI achieved a mean of .864, AMI was similar with a mean of .848.

The final solution was selected by picking the partition with the highest modularity ($Q = .428$). 31 groups were retrieved and are detailed in Table B1.

Table B1. *31 groups of subreddits retrieved by community detection algorithm.*

Subreddit group label	Example of subreddits	Number of subreddits in group
Memes/drama	r/oldpeoplefacebook, r/ShittyLifeProTips, r/iamverysmart	128
Gaming/comics	r/GlobalOffensive, r/Windows10, r/starcitizen	144
Drugs/crime	r/Drugs, r/LSD, r/Shoplifting	17
Personal/sharing	r/1200isplenty, r/ForeverAlone, r/HealthAnxiety	185
Transgender persons-related	r/lgbt, r/transpassing, r/MtF	8
Sports/politics	r/lakers, r/Europe, r/The_Donald	61
Parenting	r/beyondthebump, r/BabyBumps, r/namenerds	11
Guns	r/guns, r/ar15, r/Hunting	5
Fashion/photography/start-ups	r/malefashionadvice, r/photography, r/Entrepreneur	13
Vehicles	r/cars, r/Justrolledintotheshop, r/BMW	8
Australia	r/australia, r/brisbane, r/sydney	4
Gay male-related	r/gaybros, r/askgaybros, r/gay	3

Call of Duty	r/WWII, r/CallOfDuty	2
Paradox studio games	r/CrusaderKings, r/eu4	2
Law school	r/LawSchool, r/lawschooladmissions	2
RPGs	r/rpg, r/worldbuilding, r/dndnext	3
Graduate school	r/GradSchool, r/gradadmissions	2
Film	r/Screenwriting, r/Filmmakers	2
Investing	r/investing, r/wallstreetbets, r/stocks	3
Self-harm	r/selfharm, r/selfharpics	2
Cryptocurrency	r/litecoin, r/Bitcoin, r/ethereum	5
Magic the Gathering	r/magicTCG, r/EDH, r/mtgfinance	4
Aquariums	r/Aquariums, r/PlantedTank, r/bettafish	3
Reality TV	r/BigBrother, r/survivor	2
Programming	r/learnprogramming, r/cscareerquestions, r/webdev	3
New York City	r/nyc, r/AskNYC	2
Service staff	r/IDontWorkHereLady, r/TalesFromYourServer, r/TalesFromRetail	3
Cycling	r/cycling, r/bicycling	2
Premed	r/premed, r/Mcat	2
Nursing	r/StudentNurse, r/nursing	2
Teen Mom	r/teenmom, r/TeenMomOGandTeenMom2	2

The Personal/Sharing group was large with 185 diverse subreddits. Using the group as a single count variable would make interpretation of results difficult since the subreddits together do not have a coherent theme. Thus, text

clustering was conducted to refine this group into smaller, more topically-coherent groups. By clustering subreddits based on the words their users have used, topically-similar subreddits can be grouped together since the words appearing in them are likely to be rather alike (e.g. Sharma & De Choudhury, 2018).

Manual pruning of a number of subreddits was conducted first before proceeding with text clustering. These subreddits tended to be tangential in content (e.g. r/OrangeIsTheNewBlack, r/RupaulsDragRace). They also typically: (1) only had one link to another subreddit node within the group, and (2) that link was to one of the five very well-subscribed subreddits (redditlist, 2019) in the group ¹³(r/sex, r/askmen, r/askwomen, r/legaladvice, r/casualconversation). Removing subreddit nodes that had one degree and which were linked to one of these subreddits reduced the group size from 185 to 146. Following this, r/Canada and r/harrypotter were also removed though they had more than one link within the group as they were not of theoretical interest and were mainly linked to these five subreddits. The final count of Personal/Sharing subreddits submitted for text clustering was 144.

For each subreddit, all scraped comments written in them by the 60434 r/relationships users were concatenated into a document. In other words, there were 144 documents, each representing one subreddit and all available scraped comment content. Cleaning of text was minimal: text was lowercased for consistency and punctuation, and stop-words and Reddit usernames were removed. URL links were identified with regular expressions and converted to the string “URLLINK”. The cleaned 144 documents were then converted into a matrix of term frequency-inverse document frequency (TF-IDF) features (Pedregosa et al., 2011). TF-IDF gives greater weightage to words which are common to only a few documents (i.e., more discriminative words helpful in representing what the document is about), and lesser weightage to words which

¹³ These five subreddits also had the highest degree count and betweenness centrality (a measure of node importance) within the group.

appear very frequently across many documents (Ramos, 2003). 100000 features (unigrams and bigrams) were created, ignoring unigrams/bigrams that occurred in less than two documents (i.e., words that appeared in less than two subreddits) and unigrams/bigrams that occurred in more than 50% of the documents (i.e., words that appeared in more than 72 subreddits).

Non-negative matrix factorisation (NMF) (Pedregosa et al., 2011) was then applied onto the TF-IDF feature matrix to cluster the documents. This process followed Klein, Clutton and Polito's (2018) procedure closely. NMF requires an input of k , the number of clusters to retrieve. To assist in the selection of k , NMF was run with $k=2$ to $k=100$ in steps of 2 (i.e., $k=2, 4, 6 \dots 100$). As was done in Klein et al. (2018), the mean cosine similarity between all possible pairs of topics at each k was calculated. The mean pairwise cosine similarity for each k was then plotted out and the elbow of the plot (i.e., where the topics are largely distinct with little overlap) ($k=44$) was taken. As NMF provides soft-clustering (similar to LDA, where a document is made up of a mixture of topics), the final group membership for each subreddit was determined by taking the cluster they loaded highest in. In conclusion, 44 groups of topically-similar subreddits were created with the help of NMF from the original large Personal/Sharing group identified by the Louvain algorithm. The 44 groups were manually labelled based on the subreddits comprising them and the top 15 words identified by NMF for each group. Details of the 44 groups are given in Table B2.

Table B2. 44 groups of subreddits refined from original Personal/Sharing group.

Subreddit group label	Subreddits in group	Number of subreddits in group	Top 15 words of group
Conversation	r/AMA, r/AskAnAmerican, r/casualama, r/Christianity, r/DoesAnybodyElse, r/rant, r/self, r/SeriousConversation, r/TrueOffMyChest, r/unpopularopinion, r/WouldYouRather	11	like_butt, butt_like, black_people, christians, muslims, unpopular, christianity, islam, abortion, ama, white_people, racism, obama, feminist, muslim
Weight	r/1200isplenty, r/fatlogic, r/fatpeoplestories, r/loseit, r/proED	5	calorie, cal, kcal, bmi, fatlogic, cals, fat logic, mfp, fat_people, veggies, tdee, beetus, obesity, deficit, cico
Makeup	r/awfuleyebrows, r/Makeup, r/MakeupAddiction, r/makeupexchange, r/MakeupRehab, r/muacirclejerk, r/muacjdiscussion, r/PanPorn	8	palette, primer, eyeshadow, mascara, lipstick, matte, blush, concealer, nyx, palettes, lipsticks, sephora, eyeliner, brows, brow
Validation	r/AmIsexy, r/amiugly, r/FreeCompliments, r/Rateme	4	facial_hair, hairstyle, brows, bangs, hair_looks, eyeliner, definitely_ugly, nice_smile, beautiful_eyes, grow_hair, great_smile, first_picture, flattering, nice_eyes, rating

Food	r/AskCulinary, r/Baking, r/Cooking, r/EatCheapAndHealthy, r/MealPrepSunday	5	oven, garlic, flour, recipes, potatoes, onion, veggies, pasta, onions, tsp, olive_oil, bake, salt_pepper, seasoning, tbsp
Dating/socialising	r/AskMen, r/askseddit, r/askwomenadvice, r/dating, r/dating_advice, r/OkCupid, r/seduction, r/socialskills	8	first_date, kino, online dating, okc, escalate, second_date, flirting, okcupid, needy, seduction, ioi, fwb, make_move, nice_guy, want_date
Pets (dogs)	r/dogs, r/Dogtraining, r/pitbulls	3	leash, breed, crate, breeds, pup, pit_bull, aggression, puppies, dog_park, gsd, another_dog, pit_bulls, terrier, barking, apbt
Wedding	r/wedding, r/weddingplanning, r/Weddingsunder10k	3	fh, ceremony, bridal, venue, reception, bride, bridesmaids, moh, weddings, bridesmaid, bridal_party, wedding party, fml, photographer, rsvp
Relationship	r/BreakUps, r/ExNoContact, r/LongDistance, r/relationship_advice, r/relationships_advice	5	nc, back_together, closure, ldr, rebound, new_guy, dumper, exes, someone_new, nmom, deserve_better, broke_nc, contacting, break_nc, bu

Sex	r/AskRedditAfterDark, r/gonewild, r/sex	3	cum, orgasm, threesome, tits, clit, anal, cock, horny, vibrator, lube, panties, foreplay, condom, blowjob, masturbate
Keto	r/keto, r/ketogains, r/ketorecipes, r/xxketo	4	keto, carbs, carb, ketosis, net_carbs, macros, low_carb, tbsp, keto_friendly, heavy_cream, tsp, kcko, almond, started keto, coconut
Hair	r/curlyhair, r/FancyFollicles, r/Hair, r/malehairadvice	4	conditioner, dye, bleach, curls, curly, stylist, scalp, salon, bangs, curl, styling, frizz, comb, hair looks, toner
BDSM	r/bdsm, r/BDSMcommunity	2	dom, bdsm, submissive, kink, babygirl, safe_word, rope, fetlife, submissives, doms, domme, kinky, safeword, sadist, aftercare
Advice/help	r/Advice, r/CasualConversation, r/depression, r/offmychest, r/selfimprovement, r/socialanxiety, r/SuicideWatch	7	sa, suicidal_thoughts, therapists, loneliness, social skills, mg, want_die, antidepressants, cbt, professional_help, hotline, making_friends, lexapro, see_therapist, zoloft

Birth control	r/birthcontrol, r/childfree	2	iud, nexplanon, mirena, spotting, insertion, implant, cramps, cf, condoms, inserted, skyla, hormonal, cervix, copper_iud, uterus
Travel	r/solotravel, r/travel	2	hostel, hostels, sarajevo, flights, visa, dubrovnik, passport, tokyo, mostar, paris, tourist, touristy, prague, thailand, croatia
Skincare	r/AsianBeauty, r/SkincareAddiction	2	cleanser, bha, moisturizer, sunscreen, toner, cerave, serum, cosrx, oily, masks, skincare, cleansing, mizon, spf, ab
Running	r/C25K, r/running	2	marathon, half_marathon, runners, strava, runner, km, treadmill, keep_moving, grats, higdon, races, started_running, keep_running, hal_higdon, mileage
Pets (cats and others)	r/Blep, r/cats, r/Pets	3	kitten, kitty, litter, kittens, kitties, litter_box, blep, wet_food, rabbits, dry_food, vets, feral, spayed, older_cat, fur

Sexual orientation	r/actuallesbians, r/bisexual, r/LesbianActually	3	lesbians, bisexual, queer, bisexuals, lgbt, trans_women, bisexuality, officer, straight_people, femme, wlw, hudson, butch, trans_people, bi_people
Vegetarianism/vegan	r/vegan, r/vegetarian	2	veganism, vegans, dairy, tofu, vegetarians, eating_meat, eat_meat, cows, animal_products, veggie, soy, omni, go_vegan, ethical, vegetarianism
JustNo	r/JUSTNOFAMILY, r/JUSTNOMIL, r/JustNoSO	3	dh, mil, sil, nc, bil, fil, nmom, emotional_labor, fsil, ffil, mils, fh, lc, went_nc, time_mom
Teaching	r/Teachers, r/teaching	2	classroom, student_teaching, credential, cset, district, subbing, principal, lesson plans, admin, elementary, iep, year_teacher, certification, curriculum, assignment
Books	r/suggestmeabook, r/writing	2	fiction, plot, trilogy, genre, word_count, reader, novels, writers, protagonist, stephen, stephen_king, sci fi, short_stories, draft, published

College	r/ApplyingToCollege, r/college, r/Sat	3	gpa, scores, ap, admissions, colleges, essay, physics, ecs, scholarships, calc, tuition, majors, umass, uc, passage
Fasting	r/fasting, r/intermittentfasting	2	fasting, fasts, omad, keto, calorie, fasted, water_fast, day_fast, one_meal, meal_day, break_fast, lent, carb, broth, carbs
Piercing	r/piercing	1	piercer, piercing, piercings, pierced, jewellery, piercers, saline, helix, conch, titanium, daith, septum, barbell, nipple, tragus
Beauty gurus	r/BeautyGuruChat, r/BeautyGuruChatter	2	palette, jaclyn, tati, gurus, morphe, manny, guru, kathleen, collab, creaky, nikkie, bg, beauty_gurus, tutorials, jeffree
Confession	r/confession, r/confessions, r/DeadBedrooms, r/Marriage	4	db, spouse, intimacy, libido, hl, want_sex, sex life, resentment, affection, husbands, infidelity, marriages, orgasm, affair, sex drive
Mental health	r/ADHD, r/bipolar, r/BPD, r/mentalhealth, r/ptsd	5	bpd, ptsd, manic, mg, mania, lamictal, psychosis,

			hallucinations, hypomania, pdoc, dbt, lithium, therapists, adderall, schizoaffective
Anxiety	r/Anxiety, r/AskDocs, r/HealthAnxiety	3	health_anxiety, mri, gp, mg, brain_tumor, hiv, tumor, xanax, physician, headaches, heart_rate, diabetes, blood_test, tonsils, see_gp
Jobs	r/jobs, r/resumes	2	cover_letter, resumes, employers, entry_level, linkedin, work_experience, one_page, internship, salary, applications, gpa, template, software, verbs, employment
Tattoos	r/tattoos	1	tattoo_artist, ink, first_tattoo, shading, tattooed, sleeve, good_artist, watercolor, forearm, tat, get_tattoo, getting_tattoo, tribal, tats, love_tattoo
Fitness/bulking	r/gainit, r/progresspics, r/xxfitness	3	cardio, calorie, squats, carbs, squat, bench, workouts, bulking, tdee, lifts, reps, gains, kg, cals, macros
Incel culture	r/Incels, r/TheBluePill, r/ForeverAlone	3	trp, incel, chad, incels, twerps, normies, fa, normie, chads,

			alpha, twerp, red pill, rp, beta, rpw
Home	r/HomeImprovement, r/gardening	2	plants, soil, drywall, tile, tomatoes, sanding, grout, seeds, cabinets, carpet, hardwood, planted, install, deck, ceiling
Drinking	r/drunken, r/stopdrinking	2	drink_today, vodka, sobriety, today_drink, drinking_today, liquor, aa, whiskey, rum, hangover, fireball, hungover, quit_drinking, congrats_days, sprite
Finances/legal	r/financialindependence, r/Frugal, r/legaladvice	3	frugal, lease, fi, retirement, vanguard, fund, landlord, attorney, salary, taxes, mortgage, roth, ira, loans, loan
Parental figures	r/MomForAMinute, r/raisedbynarcissists	2	nmom, sweetie, ndad, rbn, hugs_thank, hon, sis, nparents, thankyou_much, xx, darling, sweetie, raisedbynarcissists, give_hope, going_ok
Engagement	r/JustEngaged	1	moissanite, diamond, diamonds, tiffany, sapphire, ct, jeweler, sapphires, love_ring, antique,

			engagement_ring, proposal, gibson, lab_grown, emerald
Female-centric	r/AskWomen, r/femalefashionadvice, r/indiemakeupandmore, r/RedditLaqueristas, r/TheGirlSurvivalGuide, r/TrollXChromosomes	6	boots, skirt, leather, skirts, top_coat, wardrobe, flats, leggings, opi, glitter, essie, tights, blazer, cardigan, polishes
NoFap	r/NoFap	1	nofap, pmo, fapping, fap, relapse, flatline, streak, masturbation, relapsed, urges, fapstronaut, hard_mode, dopamine, pied, masturbating
Meditation	r/Meditation	1	meditating, meditate, sensations, mindfulness, meditations, meditated, started_meditating, hallucinations, tm, guided, concentration, observe, consciousness, sleep_paralysis, mindfulness meditation
Miscellaneous	r/polyamory, r/AskWomenOver30	2	poly, polyamory, mono, monogamous, monogamy, nre, poly_people, polyamorous, okc, metamour, triad, meta, bisexual, bpd, compersion

APPENDIX C: CREATION OF VARIABLES OF STUDY.

User information. Age and gender were self-reported by the users in their post titles and/or text (e.g. I [24M] am having issues with my partner [20F]) – this is a requirement stipulated by r/relationships posting guidelines (r/relationships wiki, 2019). These were coded by hand to avoid missing out variant cases that would be difficult to catch with regular expressions (e.g. I am having issues with my [24M] partner [20F]).

Account information. Account age was calculated by taking the number of years between the date of the user's first r/relationships post in the 70520-dataset from the date of the user's first post/comment on Reddit. If the r/relationships post in the 70520-dataset happened to also be the user's first activity on Reddit, this information was captured by a separate binary variable: 1 if the post was the user's first Reddit activity, 0 if otherwise.

Activity information. The total count of all activity (posts and comments) in the year preceding the first r/relationships post in the 70520-dataset was calculated for each user. Total counts of all activity on subreddit groups of interest in the year preceding the first r/relationships post in the 70520-dataset was also calculated for each user; refer to Appendix B for subreddit group creation. The number of r/relationships comments each user wrote in the year preceding their first r/relationships post in the 70520-dataset was counted as well.

Stylistic information (post). Readability of the post was scored using the Flesch reading ease score (Flesch, 1948). The proportion of words in the post that fell under each function word category was calculated using the LIWC2015 software (Pennebaker et al., 2015). The length of post was simply the word count of the post.

Semantic information (post). The proportion of words that fell under each time focus and cognitive processes category were calculated using the LIWC2015 software (Pennebaker et al., 2015). Usage of anxiety, anger, and

sadness words was measured with the LIWC2015 software (Pennebaker et al., 2015).

Negative and positive sentiment were measured using Valence Aware Dictionary and sEntiment Reasoner (VADER), a rule-based lexicon sentiment analysis tool (Hutto & Gilbert, 2014). VADER was selected as it was created specifically for social media content and has the ability to handle negation, emoticons, slang, and intensifiers (Hutto & Gilbert, 2014). As VADER works at the sentence-level, each post was tokenised into sentences. The individual sentences were scored by VADER and the overall negative/positive sentiment of the post taken by summing the negative/positive sentiment score across all sentences divided by the number of sentences. Refer to Appendix A for topic modelling details.

Metadata. The score of the posts and the number of comments they each received were readily available from the Reddit API. Timestamps of posts and comments were also available via the API; for each post, the seconds elapsed between its publication and its first comment received were calculated by subtracting the timestamp of the post from the timestamp of the first comment.

Stylistic information (comments). Readability of the group of direct comments received by each post was scored by the Flesch reading ease score (Flesch, 1948). The proportion of words in each comment group that fell under each function word category was calculated using the LIWC2015 software (Pennebaker et al., 2015). The median length of the received direct comments for each post was calculated by taking the median word count of received direct comments per post.

For imperative identification, each comment group was tokenised into sentences and the PyDelphin library (Goodman, 2019) assisted in identifying if the sentence was imperative. The number of sentences in each comment group that were identified as imperative was tallied and divided by the total number of

sentences in the group to get the proportion of sentences in the comment group that were imperative.

For modal and hedge identification, a list of adverbs, adjectives, and phrases based on the list of hedges by Hyland (2005) was created. Each comment group was tokenised into sentences. Each word in every sentence was then tagged for its part-of-speech (e.g. noun, adverb) by the Stanford Part-Of-Speech tagger (Toutanova, Klein, Manning & Singer, 2003). A sum of modals and hedges for each comment group was created by adding the number of words that fulfilled one of these criteria:

- (1) word was tagged as a modal verb, or
- (2) word was tagged as an adverb and appeared in the prepared list, or
- (3) word was tagged as an adjective and appeared in the prepared list.

Finally, a regular expression was used to search for, in each comment group, the phrases in the prepared list. If the phrase contained two words (e.g. 'tend to'), the sum of modals and hedges was incremented by two. If the phrase contained three words, the sum was incremented by three, and so on. The proportion of modals and hedges in each comment group was calculated by dividing this sum by the word count of the comment group.

Semantic information (comments). The proportion of words that fell under each time focus and cognitive processes category were calculated using the LIWC2015 software (Pennebaker et al., 2015). Usage of anxiety, anger, and sadness words was measured with the LIWC2015 software (Pennebaker et al., 2015). Negative and positive sentiment were measured with VADER (Hutto & Gilbert, 2014) following the procedure used for the posts. Refer to Appendix A for topic modelling details.

APPENDIX D: MODEL TRAINING AND SELECTION FOR FIRST SET OF ANALYSES (REPEAT VS NON-REPEAT ADVICE-SEEKERS)

Model specifications. Using the scikit-learn library, L1-penalised logistic regression models were created with the regularisation parameter set to $C=.01$ (i.e., more regularisation) with the saga solver (Pedregosa et al., 2011). As this was a highly imbalanced task (13.2% repeat advice-seekers), the classes were weighted with weights inversely proportionate to their proportions in the training data (i.e., give more weight to the minority repeat advice-seekers class when training the model) (Pedregosa et al., 2011).

In addition, given the skewness and different ranges of the continuous variables, standardization was applied. Scikit-learn's power transformer (which applies the Yeo-Johnson transform followed by a zero-mean unit-variance standardisation) (Pedregosa et al., 2011) was selected for this task.

Model training. 5-cross-validation (5-CV) folds (i.e., shuffle the data then use 4/5 of the data for training, 1/5 for evaluating; this is done five times) on the training set were used for all models. To avoid contamination, a pipeline was set up where the power transformer was fitted on the four training folds (instead of the full training dataset) and then used to transform the held-out evaluation fold.

As with other studies that have tackled imbalanced datasets (e.g. Althoff et al., 2014; Danish, Dahiya & Talukdar, 2015; Tan et al., 2016), models were evaluated mainly using the Area Under the Receiving Operating Characteristic Curve (ROCAUC). The ROCAUC score can be understood as the probability that the model will rank a random repeat advice-seeker higher than a random non-repeat advice-seeker. Recall (calculated by $\frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})}$; the ability of the model to identify all repeat advice-seekers), precision (calculated by $\frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})}$; the ability of the model to not wrongly identify a non-repeat advice-seeker as a repeat advice-seeker), and the

F1-score (the weighted average of precision and recall) are also reported. All metrics are bound from 0 to 1, where 0 indicates the worst performance and 1 indicates the best. For the ROCAUC score, a random classifier typically performs at 0.5.

Scikit-learn's dummy classifier (Pedregosa et al., 2011) was used as a baseline; the classifier outputs predictions according to the class proportions of the training folds. The variable sets were first evaluated individually (Table D1). Following this, the variable sets were added incrementally to the model based on their individual predictive utility (Table D2; selected model is bolded). If the variable set did not improve model performance, it was dropped from further consideration.

Table D1. *Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – base model.*

Model	Precision	Recall	F1	ROC AUC
Dummy classifier	.142	.157	.149	.505
User information	.155	.564	.243	.558
Account information	.158	.633	.253	.582
Activity information	.174	.354	.233	.571
Post stylistic information	.152	.509	.234	.547
Post semantic information (LDA)	.153	.543	.238	.552
Post semantic information (Time focus)	.141	.509	.221	.523
Post semantic information (Cognitive processes)	.139	.493	.217	.507
Post semantic information (Sentiment)	.139	.516	.219	.513
Post metadata information	.136	.522	.216	.506
Comment stylistic information	.135	.468	.210	.498
Comment semantic information (LDA)	.145	.545	.228	.535

Comment semantic information (Time focus)	.135	.481	.211	.501
Comment semantic information (Cognitive processes)	.141	.501	.219	.515
Comment semantic information (Sentiment)	.134	.482	.209	.503

Table D2. *Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – base model.*

Model	Precision	Recall	F1	ROC AUC
Dummy classifier	.142	.157	.149	.505
Full model (all variable sets)	.177	.597	.273	.620
User + Account	.160	.622	.255	.587
User + Account + Activity	.166	.646	.265	.611
User + Account + Activity + Post LDA	.175	.6	.271	.619
User + Account + Activity + Post LDA + Post Stylistic	.178	.603	.275	.621
User + Account + Activity + Post LDA + Post Stylistic + Comment LDA	.177	.601	.273	.620
User + Account + Activity + Post LDA + Post Stylistic + Post Time Focus	.177	.602	.274	.620
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes	.178	.604	.275	.621
User + Account + Activity + Post LDA + Post Stylistic + Post Sentiment	.176	.597	.272	.622
User + Account + Activity + Post LDA + Post Stylistic + Post Metadata	.177	.602	.274	.622
User + Account + Activity + Post LDA + Post Stylistic + Comments Stylistic	.178	.601	.275	.621
User + Account + Activity + Post LDA + Post Stylistic + Comments Time Focus	.177	.601	.274	.621

User + Account + Activity + Post LDA + Post Stylistic + Comments Cognitive Processes	.178	.602	.274	.621
User + Account + Activity + Post LDA + Post Stylistic + Comments Sentiment	.178	.602	.274	.620

Model testing. The selected model (user + account + activity + post LDA + post stylistic) was trained on the full training dataset and applied to the testing set. Its performance was ROCAUC = .587, precision = .170, recall = .594, F1 = .264.

Non-zero coefficients of final model. Table D3 displays the non-zero coefficients of the final model. Note that due to the Yeo-Johnson transformation, the magnitude of the coefficients may not be directly comparable with each other; however, the direction of the coefficients can be interpreted given that it is a monotonic transformation.

Table D3. *Non-zero coefficients (base model).*

Intercept	-.119
Age of user	-.048
Account age	-.287
Total count of account activity in preceding year	.014
Prior comments on r/relationships in preceding year	.146
Activity on dating-related subreddits in preceding year	.044
Activity on wedding-related subreddits in preceding year	-.050
Activity on relationship-related subreddits in preceding year	.079
Activity on advice/help-related subreddits in preceding year	.045
Activity on sexual orientation-related subreddits in preceding year	-.033
Activity on confession-related subreddits in preceding year	.005
Activity on anxiety-related subreddits in preceding year	.072
Activity on meditation-related subreddits in preceding year	.019
Body image (post LDA)	-.028

Communication (post LDA)	.024
Dating (post LDA)	.027
Family (post LDA)	-.012
Jobs (post LDA)	-.013
Marital (post LDA)	-.066
Mental health (post LDA)	-.034
Past partners (post LDA)	.012
Romantic partners (post LDA)	.012
School (post LDA)	-.007
Sexual (post LDA)	.010
Texting (post LDA)	.099
Trips/occasions (postLDA)	-.005
Length of post	-.058
Usage of first-person singular pronouns (e.g. 'I')	.037
Usage of third-person singular pronouns (e.g. 'she')	.009
Usage of prepositions	-.001
Usage of conjunctions	.042
Gender of user	.092

**APPENDIX E: RESULTS OF KRUSKAL-WALLIS TESTS ON
REPEAT/NON-REPEAT ADVICE-SEEKERS SPLIT BY GENDER.**

Table E1 displays the significant results of the Kruskal-Wallis tests with Holm-Bonferroni correction ($\alpha = .05$) on all 66 variables used in training the model from the first set of analyses. The four groups compared were male/female non-repeat advice-seekers and male/female repeat advice-seekers. Only data used in training the model in the first set of analyses was used for this comparison.

Table E1. *Significant results of Kruskal-Wallis tests after Holm-Bonferroni correction.*

(MNR = male non-repeat advice-seeker, FNR = female non-repeat advice-seeker, MR = male repeat advice-seeker, FR = female repeat advice-seeker)

(Sig = significant, Non-sig = non-significant)

Variable	Cohen's d	Pairwise differences	Mean ranks
Self-reported age	.176***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p = .006$ MR-FNR: $p < .001$ FNR-FR: $p < .001$ Non-sig: MNR-FR: $p = 1.000$ MR-FR: $p = .169$	MR=6429.16 FR=6845.07 MNR=6921.14 FNR=7568.51
Account age at time of first r/relationships post in 70520-dataset	.336***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p < .001$ MNR-FR: $p < .001$ MR-FR: $p < .001$ FNR-FR: $p < .001$ Non-sig: MR-FNR: $p = 1.000$	FR = 5799.34 FNR = 6750.69 MR = 6931.19 MNR = 7801.91

LDA post topic: Communication	.201***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p = .004$ MR-FR: $p < .001$ Non-sig: MNR-MR: $p = .431$ FNR-FR: $p = .081$	MNR = 6725.68 MR = 6997.12 FNR = 7515.91 FR = 7855.12
LDA post topic: Dating	.352***	MNR-FNR: $p < .001$ MNR-MR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p < .001$ FNR-FR: $p < .001$	FNR = 6372.94 FR = 6943.32 MNR = 7780.90 MR = 8324.31
LDA post topic: Drinking/drugs/ religion	.097***	Sig: MNR-FNR: $p < .001$	FNR = 6921.18 MR = 7207.99 FR = 7219.61 MNR = 7365.86
LDA post topic: Family	.16***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p = .046$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p < .001$ Non-sig: FNR-FR: $p = .663$	MR = 6469.50 MNR = 6891.03 FR = 7293.32 FNR = 7512.35

LDA post topic: Fighting/arguments	.189***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p = .003$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	MR = 6726.42 MNR = 6778.92 FR = 7390.63 FNR = 7580.23
LDA post topic: Finances	.161***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p < .001$ MNR-FR: $p = .014$ MR-FNR: $p < .001$ MR-FR: $p < .001$ Non-sig: FNR-FR: $p = 1.000$	MR = 6330.15 MNR = 6926.72 FR = 7342.63 FNR = 7488.37
LDA post topic: Friends	.131***	Sig: MNR-FNR: $p < .001$ MR-FNR: $p = .001$ FNR-FR: $p = .041$ Non-sig: MNR-MR: $p = 1.000$ MNR-FR: $p = .769$ MR-FR: $p = 1.000$	FNR = 6838.95 FR = 7210.40 MR = 7408.85 MNR = 7418.25

LDA post topic: Getting married	.058**	Sig: MNR-FNR: p=.005	FR = 7012.64 FNR = 7027.59 MNR = 7276.08 MR = 7386.96
LDA post topic: Hobbies/interests	.066***	Sig: MNR-FNR: p<.001	FNR = 6982.91 MR = 7269.20 MNR = 7286.89 FR = 7294.00
LDA post topic: Long-distance relationships	.563***	Sig: MNR-FNR: p<.001 MNR-FR: p<.001 MR-FNR: p<.001 MR-FR: p<.001 Non-sig: MNR-MR: p=1.000 FNR-FR: p=.937	FNR = 5999.45 FR = 6194.04 MNR = 8252.89 MR = 8372.30
LDA post topic: Living arrangements (hygiene/chores)	.194***	Sig: MNR-FNR: p<.001 MNR-FR: p=.017 MR-FNR: p<.001 MR-FR: p<.001 FNR-FR: p=.041 Non-sig: MNR-MR: p=.095	MR = 6458.58 MNR = 6822.28 FR = 7229.60 FNR = 7600.79

LDA post topic: Marital	.12***	Sig: MNR-FNR: $p < .001$ FNR-MR: $p < .001$	MR = 6606.25 MNR = 6982.21 FR = 7091.02 FNR = 7435.20
LDA post topic: Mental Health	.187***	Sig: MNR-MR: $p = .040$ MNR-FNR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p = .011$ FNR-FR: $p < .001$ Non-sig: MNR-FR: $p = 1.000$	MR = 6450.11 MNR = 6858.73 FR = 7042.41 FNR = 7596.60
LDA post topic: Moving	.102***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p = .015$ MR-FNR: $p < .001$ MR-FR: $p = .001$ Non-sig: MNR-FR: $p = .318$ FNR-FR: $p = 1.000$	MR = 6568.16 MNR = 7022.89 FR = 7287.22 FNR = 7362.73

LDA post topic: Past partners	.201***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .002$ MR-FNR: $p < .001$ MR-FR: $p < .001$ Non-sig: MNR-MR: $p = .395$ FNR-FR: $p = .074$	FNR = 6699.13 FR = 7042.45 MNR = 7525.78 MR = 7803.12
LDA post topic: Romantic feelings	.155***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .001$ MR-FNR: $p < .001$ MR-FR: $p = .001$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	FNR = 6817.03 FR = 6929.22 MNR = 7455.78 MR = 7635.41
LDA post topic: School	.123***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .001$	FNR = 6894.77 FR = 6920.59 MR = 7284.27 MNR = 7430.72

LDA post topic: Sexual	.185***	Sig: MNR-FNR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p = .003$ FNR-FR: $p = .007$ Non-sig: MNR-FR: $p = .173$ MNR-MR: $p = .105$	FNR = 6728.82 FR = 7173.31 MNR = 7472.13 MR = 7830.54
LDA post topic: Texting	.147***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p = .020$ FNR-FR: $p < .001$ Non-sig: MNR-FR: $p = 1.000$	FNR = 6847.05 MNR = 7298.88 FR = 7422.03 MR = 7979.30
Readability of post	.083***	Sig: MNR-FNR: $p = .002$ MNR-FR: $p < .001$	MNR = 6982.63 MR = 7247.36 FNR = 7252.46 FR = 7607.44

Length of post	.198***	Sig: MNR-FNR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p = .016$ FNR-FR: $p < .001$ Non-sig: MNR-MR: $p = .127$ MNR-FR: $p = .631$	MR = 6479.20 MNR = 6826.55 FR = 7047.99 FNR = 7625.56
First-person singular pronoun usage in post (e.g. 'I')	.136***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FR: $p = .031$ Non-sig: MNR-MR: $p = .276$ MR-FNR: $p = .846$ FNR-FR: $p = .148$	MNR = 6858.36 MR = 7159.17 FNR = 7381.82 FR = 7690.04
First-person plural pronoun usage in post (e.g. 'we')	.093***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .006$	FR = 6905.96 FNR = 6964.46 MR = 7354.03 MNR = 7358.00

Third-person plural pronoun usage in post (e.g. 'they')	.164***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p < .001$ Non-sig: MNR-MR: $p = .228$ FNR-FR: $p = 1.000$	MR = 6575.34 MNR = 6875.84 FR = 7451.97 FNR = 7488.14
Impersonal pronoun usage in post (e.g. 'it')	.134***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .010$ MR-FNR: $p < .001$ Non-sig: MNR-MR: $p = 1.000$ MR-FR: $p = .109$ FNR-FR: $p = 1.000$	FNR = 6825.43 FR = 7005.85 MNR = 7433.69 MR = 7454.39
Function word usage (article) in post	.135***	Sig: MNR-FNR: $p < .0001$ MNR-FR: $p < .001$	FR = 6772.88 FNR = 6905.94 MR = 7093.13 MNR = 7470.46
Function word usage (preposition) in post	.15***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p = .018$ MNR-FR: $p < .001$ Non-sig: FNR-MR: $p = 1.000$ FNR-FR: $p = .998$	FR = 6701.70 FNR = 6891.61 MR = 7053.32 MNR = 7501.24

Function word usage (auxiliary verb) in post	.139***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p = .010$ MR-FR: $p = .034$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	MNR = 6856.94 MR = 6959.29 FNR = 7444.93 FR = 7494.14
Function word usage (conjunction) in post	.136***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p = .001$ MNR-FR: $p < .001$ FNR-FR: $p = .040$ Non-sig: MR-FNR: $p = 1.000$ MR-FR: $p = .928$	MNR = 6856.13 FNR = 7340.18 MR = 7442.19 FR = 7712.35
Function word usage (negation) in post	.157***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p = .044$ MR-FR: $p = .014$ Non-sig: MNR-MR: $p = .668$ FNR-FR: $p = 1.000$	MNR = 6811.42 MR = 7051.47 FNR = 7457.18 FR = 7631.24

Comments on r/relationships for one year prior to first post in 70520-dataset	.301***	MNR-FNR: $p < .001$ MNR-MR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p = .001$ MR-FR: $p < .001$ FNR-FR: $p < .001$	MNR = 6756.82 MR = 7102.54 FNR = 7465.37 FR = 7870.10
Total account activity on Reddit for one year prior to first post in 70520-dataset	.196***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p < .001$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	FNR = 6742.89 FR = 6800.44 MNR = 7556.00 MR = 7571.14
Activity on drug/crime-related subreddits for one year prior to first post in 70520-dataset	.12***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .002$ MR-FNR: $p = .001$ MR-FR: $p = .016$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	FNR = 7046.31 FR = 7066.14 MNR = 7247.33 MR = 7275.52

Activity on parenting-related subreddits for one year prior to first post in 70520-dataset	.173***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .001$ MR-FNR: $p < .001$ MR-FR: $p = .012$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = .668$	MR = 7064.41 MNR = 7079.12 FR = 7198.65 FNR = 7248.71
Activity on gay male-related subreddits for one year prior to first post in 70520-dataset	.125***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .001$ MR-FNR: $p = .006$ Non-sig: MNR-MR: $p = 1.000$ MR-FR: $p = .079$ FNR-FR: $p = 1.000$	FNR = 7116.41 FR = 7121.00 MR = 7193.83 MNR = 7199.82
Activity on self-harm-related subreddits for one year prior to first post in 70520-dataset	.06**	Sig: MNR-FNR: $p = .002$	MR = 7142.50 MNR = 7148.12 FR = 7169.30 FNR = 7169.84

Activity on weight-related subreddits for one year prior to first post in 70520-dataset	.177***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p = .017$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	MNR = 7043.41 MR = 7062.27 FR = 7237.56 FNR = 7279.91
Activity on dating-related subreddits for one year prior to first post in 70520-dataset	.119***	Sig: MNR-FNR: $p < .001$ MNR-MR: $p < .001$ FNR-MR: $p < .001$ FNR-FR: $p = .006$ Non-sig: MNR-FR: $p = 1.000$ MR-FR: $p = .093$	FNR = 7032.44 MNR = 7221.03 FR = 7248.52 MR = 7468.90
Activity on wedding-related subreddits for one year prior to first post in 70520-dataset	.172***	Sig: MNR-FNR: $p < .001$ MR-FNR: $p < .001$ FNR-FR: $p < .001$ Non-sig: MNR-MR: $p = 1.000$ MNR-FR: $p = 1.000$ MR-FR: $p = 1.000$	MR = 7104.50 MNR = 7110.10 FR = 7138.02 FNR = 7221.01

Activity on relationship-related subreddits for one year prior to first post in 70520-dataset	.118***	Sig: MNR-FR: $p < .001$ FNR-FR: $p < .001$ MR-FR: $p = .009$ Non-sig: MNR-MR: $p = .192$ MNR-FNR: $p = 1.000$ MR-FNR: $p = .151$	FNR = 7109.68 MNR = 7117.40 MR = 7280.10 FR = 7582.10
Activity on advice/help-related subreddits for one year prior to first post in 70520-dataset	.07***	Sig: MNR-MR: $p = .033$ MNR-FR: $p = .001$ FNR-FR: $p = .008$ Non-sig: MNR-FNR: $p = 1.000$ MR-FNR: $p = .196$ MR-FR: $p = 1.000$	MNR = 7098.61 FNR = 7153.01 MR = 7337.37 FR = 7404.19
Activity on sexual orientation-related subreddits for one year prior to first post in 70520-dataset	.127***	Sig: MNR-FNR: $p < .001$ MR-FNR: $p = .001$	MR = 7107.83 MNR = 7115.02 FR = 7149.27 FNR = 7213.33

Activity on justno-themed subreddits for one year prior to first post in 70520-dataset	.154***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p = .014$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	MR = 7097.44 MNR = 7101.80 FR = 7201.13 FNR = 7219.59
Activity on confession-related subreddits for one year prior to first post in 70520-dataset	.086***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$	MNR = 7104.66 MR = 7145.67 FNR = 7197.58 FR = 7269.56
Activity on mental health-related subreddits for one year prior to first post in 70520-dataset	.097***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p = .034$	MNR = 7099.67 MR = 7154.39 FR = 7197.18 FNR = 7213.07
Activity on anxiety-related subreddits for one year prior to first post in 70520-dataset	.101***	Sig: MNR-FNR: $p = .003$ MNR-MR: $p = .006$ MNR-FR: $p < .001$ FNR-FR: $p = .002$ Non-sig: FNR-MR: $p = .701$ MR-FR: $p = 1.000$	MNR = 7097.50 FNR = 7178.63 MR = 7252.97 FR = 7333.69

Activity on parental figure-related subreddits for one year prior to first post in 70520-dataset	.16***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p = .001$ MR-FR: $p < .001$ FNR-FR: $p = .048$ Non-sig: MNR-MR: $p = 1.000$	MNR = 7079.26 MR = 7087.62 FNR = 7224.81 FR = 7314.02
Activity on NoFap subreddit for one year prior to first post in 70520-dataset	.208***	Sig: MNR-FNR: $p < .001$ MNR-FR: $p < .001$ MR-FNR: $p < .001$ MR-FR: $p < .001$ Non-sig: MNR-MR: $p = 1.000$ FNR-FR: $p = 1.000$	FNR = 7073.30 FR = 7071.64 MNR = 7239.08 MR = 7266.78

** $p < .01$, *** $p < .001$

APPENDIX F: MODEL TRAINING AND SELECTION FOR SECOND SET OF ANALYSES (MALE REPEAT VS MALE NON-REPEAT ADVICE-SEEKERS).

Model specifications. As done in the first set of analyses (Appendix D), L1-penalised logistic regression models were created with the regularisation parameter set to $C=.01$ with the saga solver (Pedregosa et al., 2011). As this was also a highly imbalanced task (11.6% repeat advice-seekers), the classes were weighted with weights inversely proportionate to their proportions in the training data (Pedregosa et al., 2011). Continuous variables were transformed with scikit-learn's power transformer (the Yeo-Johnson transform followed by a zero-mean unit-variance standardisation) (Pedregosa et al., 2011).

Model training. 5-CV folds on the training set were used for all models. A pipeline was set up where the power transformer was fitted on the four training folds (instead of the full training dataset) and then used to transform the held-out evaluation fold. Models were again evaluated mainly using ROCAUC. Recall, precision and the F1-score are also reported. Scikit-learn's dummy classifier (Pedregosa et al., 2011) was used as a baseline; it output predictions according to the class proportions of the training folds. The variable sets were first evaluated individually (Table F1), then incrementally based on their individual predictive utility (Table F2; selected model is bolded). If the variable set did not improve model performance, it was dropped from further consideration.

Table F1. *Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – male-only model.*

Model	Precision	Recall	F1	ROC AUC
Dummy classifier	.112	.126	.119	.497
User information (Age)	.131	.525	.210	.534
Account information	.132	.542	.213	.551

Activity information	.154	.251	.191	.536
Post stylistic information	.132	.521	.211	.548
Post semantic information (LDA)	.128	.554	.208	.539
Post semantic information (Time focus)	.098	.411	.158	.516
Post semantic information (Cognitive processes)	.069	.383	.115	.496
Post semantic information (Sentiment)	.126	.545	.205	.527
Post metadata information	.069	.402	.116	.494
Comment stylistic information	.127	.482	.201	.525
Comment semantic information (LDA)	.121	.542	.197	.524
Comment semantic information (Time focus)	.073	.398	.121	.503
Comment semantic information (Cognitive processes)	.072	.418	.121	.502
Comment semantic information (Sentiment)	.069	.288	.112	.492

Table F2. *Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – male-only model.*

Model	Precision	Recall	F1	ROC AUC
Dummy classifier	.112	.126	.119	.497
Full model (all variable sets)	.141	.541	.224	.577
User (Age) + Account	.131	.549	.212	.561
User (Age) + Account + Activity	.133	.614	.219	.574
User (Age) + Account + Activity + Post LDA	.137	.547	.219	.577
User (Age) + Account + Activity + Post LDA + Post Stylistic	.140	.543	.223	.578
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment	.142	.546	.226	.580

User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Comments Stylistic	.143	.547	.227	.580
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Comment LDA	.141	.543	.224	.577
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Post Time Focus	.142	.545	.225	.580
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Post Cognitive Processes	.142	.545	.225	.580
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Post Metadata	.143	.549	.226	.580
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Comments Time Focus	.142	.545	.225	.580
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Comments Cognitive Processes	.142	.547	.226	.580
User (Age) + Account + Activity + Post LDA + Post Stylistic + Post Sentiment + Comments Sentiment	.142	.542	.224	.579

Model testing. The selected model (user age + account + activity + post LDA + post stylistic + post sentiment) was trained on the full training dataset and applied to the testing set. Its performance was ROCAUC = .588, precision = .151, recall = .597, F1 = .240.

Non-zero coefficients of final model. Table F3 displays the non-zero coefficients of the final model. Again, due to the Yeo-Johnson transformation,

the magnitude of the coefficients may not be directly comparable with each other; however, the direction of the coefficients can be interpreted.

Table F3. *Non-zero coefficients (male-only model).*

Intercept	-.040
Age of user	-.001
Account age	-.159
Total count of account activity in preceding year	.029
Prior comments on r/relationships in preceding year	.089
Activity on dating-related subreddits in preceding year	.002
Activity on advice/help-related subreddits in preceding year	.001
Activity on mental health-related subreddits in preceding year	.005
Activity on anxiety-related subreddits in preceding year	.062
Dating (post LDA)	.024
Drinking/drugs/religion (post LDA)	-.034
Finances (post LDA)	-.061
Jobs (post LDA)	-.028
Marital (post LDA)	-.030
Social media/trust issues (post LDA)	.003
Texting (post LDA)	.039
Length of post	-.007
Usage of first-person singular pronouns (e.g. 'I')	.012
Usage of third-person singular pronouns (e.g. 'she')	.029
Usage of articles	-.006
Usage of prepositions	-.077
Usage of conjunctions	.014
Use of anxiety words	.045

APPENDIX G: MODEL TRAINING AND SELECTION FOR SECOND SET OF ANALYSES (FEMALE REPEAT VS FEMALE NON-REPEAT ADVICE-SEEKERS)/

Model specifications. As done in the first set of analyses (Appendix D), L1-penalised logistic regression models were created with the regularisation parameter set to $C=.01$ with the saga solver (Pedregosa et al., 2011). As this was also a highly imbalanced task (15% repeat advice-seekers), the classes were weighted with weights inversely proportionate to their proportions in the training data (Pedregosa et al., 2011). Continuous variables were transformed with scikit-learn's power transformer (the Yeo-Johnson transform followed by a zero-mean unit-variance standardisation) (Pedregosa et al., 2011).

Model training. 5-CV folds on the training set were used for all models. A pipeline was set up where the power transformer was fitted on the four training folds (instead of the full training dataset) and then used to transform the held-out evaluation fold. Models were again evaluated mainly using ROCAUC. Recall, precision and the F1-score are also reported. Scikit-learn's dummy classifier (Pedregosa et al., 2011) was used as a baseline; it output predictions according to the class proportions of the training folds. The variable sets were first evaluated individually (Table G1), then incrementally based on their individual predictive utility (Table G2; selected model is bolded). If the variable set did not improve model performance, it was dropped from further consideration.

Table G1. *Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – female-only model.*

Model	Precision	Recall	F1	ROC AUC
Dummy classifier	.157	.177	.167	.505
User information (Age)	.177	.574	.270	.558
Account information	.173	.690	.277	.565

Activity information	.195	.379	.257	.556
Post stylistic information	.159	.476	.238	.522
Post semantic information (LDA)	.178	.568	.271	.559
Post semantic information (Time focus)	.161	.534	.248	.530
Post semantic information (Cognitive processes)	.158	.510	.241	.530
Post semantic information (Sentiment)	.119	.700	.201	.496
Post metadata information	.118	.509	.189	.492
Comment stylistic information	.153	.484	.233	.515
Comment semantic information (LDA)	.156	.546	.242	.525
Comment semantic information (Time focus)	.091	.382	.143	.502
Comment semantic information (Cognitive processes)	.158	.532	.244	.517
Comment semantic information (Sentiment)	.122	.489	.192	.506

Table G2. *Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – female-only model.*

Model	Precision	Recall	F1	ROC AUC
Dummy classifier	.157	.177	.167	.505
Full model (all variable sets)	.200	.611	.301	.622
User (Age) + Account	.179	.639	.279	.586
User (Age) + Account + Activity	.197	.619	.298	.619
User (Age) + Account + Activity + Post LDA	.201	.630	.305	.626
User (Age) + Account + Activity + Post LDA + Post Stylistic	.198	.627	.301	.621
User (Age) + Account + Activity + Post LDA + Post Time Focus	.200	.623	.302	.627

User (Age) + Account + Activity + Post LDA + Post Cognitive Processes	.200	.628	.304	.626
User (Age) + Account + Activity + Post LDA + Comment LDA	.195	.646	.300	.619
User (Age) + Account + Activity + Post LDA + Post Sentiment	.200	.628	.303	.626
User (Age) + Account + Activity + Post LDA + Post Metadata	.201	.630	.305	.626
User (Age) + Account + Activity + Post LDA + Comments Stylistic	.201	.631	.305	.627
User (Age) + Account + Activity + Post LDA + Comments Time Focus	.201	.630	.305	.626
User (Age) + Account + Activity + Post LDA + Comments Cognitive Processes	.202	.629	.306	.628
User (Age) + Account + Activity + Post LDA + Comments Cognitive Processes + Comments Sentiment	.203	.630	.307	.628

Model testing. The selected model (user age + account + activity + post LDA + cognitive processes words in received comments) was trained on the full training dataset and applied to the testing set. Its performance was ROCAUC = .604, precision = .204, recall = .649, F1 = .310.

Non-zero coefficients of final model. Table G3 displays the non-zero coefficients of the final model. Again, due to the Yeo-Johnson transformation, the magnitude of the coefficients may not be directly comparable with each other; however, the direction of the coefficients can be interpreted.

Table G3. *Non-zero coefficients (female-only model).*

Intercept	-.059
Age of user	-.123
Account age	-.252
Total count of account activity in preceding year	.004

Prior comments on r/relationships in preceding year	.130
Activity on dating-related subreddits in preceding year	.034
Activity on relationship-related subreddits in preceding year	.101
Activity on advice/help-related subreddits in preceding year	.024
Activity on sexual orientation-related subreddits in preceding year	-.015
Activity on anxiety-related subreddits in preceding year	.054
Activity on parental figure-related subreddits in preceding year	.008
Communication (post LDA)	.037
Marital (post LDA)	-.035
Mental health (post LDA)	-.001
Sexual (post LDA)	.027
Texting (post LDA)	.068
Trips/occasions (post LDA)	-.001
Causation words in comments	.041

APPENDIX H: MODEL TRAINING AND SELECTION FOR THIRD SET OF ANALYSES (ACCOUNTING FOR ADVICE-GIVERS).

Model specifications. L2-penalised multinomial logistic regression models were created with the regularisation parameter set to $C=.01$ (i.e., more regularisation) with the lbfgs solver (Pedregosa et al., 2011). As the classes were imbalanced, the classes were weighted with weights inversely proportionate to their proportions in training data (Pedregosa et al., 2011). Continuous variables were transformed with scikit-learn's power transformer (the Yeo-Johnson transform followed by a zero-mean unit-variance standardisation) (Pedregosa et al., 2011).

Model training. 5-CV folds on the training set were used for all models. A pipeline was set up where the power transformer was fitted on the four training folds (instead of the full training dataset) and then used to transform the held-out evaluation fold.

Given that this was a multiclass problem, different evaluation metrics were used from the first two sets of analyses. Cohen's kappa (Cohen, 1960) – a coefficient that conveys the inter-annotator agreement and that corrects for chance agreement – was used. Macro-averaged precision, recall, and F1 scores were calculated. Macro-averaging simply sums the scores (e.g. precision) for each individual class and divides it by the number of classes (in this case, four). It thus weighs each class equally – the smaller classes are given as much weight as the majority complete non-repeat class. For comparison, the micro-F1 score was also calculated. Micro-averaging takes into account class sizes. For example, in the case of the precision score, a micro-average would sum the count of true-positives across all classes, sum the count of true-positives and false-positives of each class across all four classes, and then divide the former by the latter. For all metrics, a higher score indicates better performance.

Scikit-learn's dummy classifier (Pedregosa et al., 2011) was used as a baseline; it output predictions according to the class proportions of the training

folds. The variable sets were first evaluated individually (Table H1), then incrementally based on their individual predictive utility (Table H2; selected model is bolded). If the variable set did not improve model performance, it was dropped from further consideration.

Table H1. *Individual variable sets (averaged performance on evaluation fold of 5-CV on train set) – multinomial model.*

Model	Cohen's Kappa	Precision Macro	Recall Macro	F1 Macro	F1 Micro
Dummy classifier	.004	.255	.256	.255	.644
User information	.043	.279	.328	.223	.320
Account information	.017	.063	.319	.102	.083
Activity information	.088	.320	.354	.291	.455
Post stylistic information	.023	.267	.289	.189	.259
Post semantic information (LDA)	.023	.266	.291	.176	.226
Post semantic information (Time focus)	.016	.230	.281	.167	.221
Post semantic information (Cognitive processes)	.014	.261	.268	.184	.264
Post semantic information (Sentiment)	.004	.249	.265	.127	.139
Post metadata information	.015	.262	.267	.184	.263
Comment stylistic information	.017	.265	.278	.165	.210
Comment semantic information (LDA)	.017	.264	.280	.168	.212
Comment semantic information (Time focus)	.004	.257	.248	.125	.143
Comment semantic information (Cognitive processes)	$2.97e^{-05}$	.247	.251	.116	.122

Comment semantic information (Sentiment)	-.003	.248	.245	.146	.187
--	-------	------	------	------	------

Table H2. *Incremental variable sets (averaged performance on evaluation fold of 5-CV on train set) – multinomial model.*

Model	Cohen's Kappa	Precision Macro	Recall Macro	F1 Macro	F1 Micro
Dummy classifier	.004	.255	.256	.255	.644
Full model	.090	.306	.355	.273	.419
User + Account	.048	.288	.344	.209	.269
User + Account + Activity	.101	.322	.379	.290	.426
User + Account + Activity + Post LDA	.110	.319	.376	.294	.452
User + Account + Activity + Post LDA + Post Stylistic	.112	.320	.384	.296	.448
User + Account + Activity + Post LDA + Post Stylistic + Post Time Focus	.112	.320	.385	.296	.449
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes	.115	.322	.387	.298	.452
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes + Post Sentiment	.110	.319	.381	.295	.449
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes + Comment LDA	.096	.310	.365	.281	.429
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive	.111	.320	.385	.295	.447

Processes + Post Metadata					
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes + Comments Stylistic	.108	.318	.380	.292	.443
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes + Comments Time Focus	.113	.321	.385	.297	.450
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes + Comments Cognitive Processes	.111	.320	.382	.295	.448
User + Account + Activity + Post LDA + Post Stylistic + Post Cognitive Processes + Comments Sentiment	.111	.319	.382	.295	.449

Model testing. The selected model (user + account + activity + post LDA + post stylistic + cognitive processes word usage in post) was trained on the full training dataset and applied to the testing set. Its performance was Cohen's kappa = .111, macro-precision = .321, macro-recall = .370, macro-F1 = .292, micro-F1 = .432.

Model coefficients. Table H3 displays the coefficients of all four equations derived from the multinomial logistic regression model.

Table H3. *Coefficients from multinomial logistic regression model.*

	Full non-return	Only advice-giving	Only advice-seeking	Full return
Intercept	.412	-.340	.141	-.213
Gender	-.272	.248	-.166	.189
User age	-.033	.193	-.162	.002
Account age	.255	.047	-.025	-.276
Prior comments on r/relationships in preceding year	-.332	.243	-.113	.202
Total count of account activity in preceding year	-.240	.315	-.194	.118
Activity on drug/crime-related subreddits in preceding year	.027	-.042	.063	-.047
Activity on transgender persons-related subreddits in preceding year	-.001	.022	-.004	-.017
Activity on parenting-related subreddits in preceding year	.035	-.007	-.120	.092
Activity on gay men-related subreddits in preceding year	-.002	-.048	.031	.019
Activity on self-harm-related subreddits in preceding year	.021	.038	-.096	.037
Activity on weight-related subreddits in preceding year	-.012	.005	-.007	.015
Activity on validation-related subreddits in preceding year	.028	.014	-.024	-.018
Activity on dating-related subreddits in preceding year	-.025	-.029	.053	.001
Activity on wedding-related subreddits in preceding year	.045	.049	-.051	-.043
Activity on relationship-related subreddits in preceding year	-.043	-.073	.069	.046
Activity on sex-related subreddits in preceding year	-.033	.026	-.013	.020
Activity on advice/help-related subreddits in preceding year	.013	-.047	.069	-.035
Activity on sexual orientation-related subreddits in preceding year	.028	.038	-.038	-.028
Activity on JustNo-themed subreddits in preceding year	.018	.014	.012	-.043

Activity on confession-related subreddits in preceding year	-.036	-.001	.047	-.010
Activity on mental health-related subreddits in preceding year	-.009	-.024	.008	.025
Activity on anxiety-related subreddits in preceding year	-.042	-.005	.029	.018
Activity on incel-related subreddits in preceding year	-.034	.012	.032	-.011
Activity on drinking-related subreddits in preceding year	-.011	.005	-.037	.043
Activity on parental figure-related subreddits in preceding year	-.026	-.017	.029	.072
Activity on engagement subreddit in preceding year	.014	-.033	-.040	.060
Activity on NoFap subreddit in preceding year	.003	.032	-.035	.001
Activity on meditation subreddit in preceding year	-.023	.000	.062	-.039
Post LDA: Body image	.020	.029	-.027	-.021
Post LDA: Communication	-.026	-.037	.069	-.005
Post LDA: Dating	-.004	.012	.052	-.060
Post LDA: Drinking/drugs/religion	.039	-.024	-.007	-.008
Post LDA: Family	.045	-.001	-.003	-.041
Post LDA: Fighting/arguments	-.009	.003	-.025	.030
Post LDA: Finances	.001	-.037	.001	.035
Post LDA: Friends	-.003	.037	-.021	-.012
Post LDA: Getting married	.006	.026	-.033	.002
Post LDA: Hobbies/interests	.042	-.039	.027	-.030
Post LDA: Jobs	0.06	-.036	.023	-.044
Post LDA: LDR	-.042	.009	-.006	.040
Post LDA: Living arrangements (hygiene/chores)	-.026	.053	-.005	-.022
Post LDA: Living arrangements (space)	.016	-.021	-.033	.038
Post LDA: Marital	.038	.017	-.063	.008
Post LDA: Mental health	.054	.000	.026	-.080
Post LDA: Moving	-.015	.060	-.006	-.039
Post LDA: Past partners	-.031	-.025	.001	.054
Post LDA: Romantic feelings	-.041	.037	-.028	.032
Post LDA: School	.029	.018	-.042	-.006
Post LDA: Sexual	.006	-.052	.021	.025

Post LDA: Social media/trust issues	.009	-.006	.002	-.004
Post LDA: Socializing	.010	-.026	-.008	.024
Post LDA: Texting	-.020	-.067	.118	-.031
Post LDA: Trips/occasions	.004	.060	.011	-.074
Readability of post	.008	-.059	.028	.023
Length of post	-.025	.119	-.059	-.035
Usage of first-person singular pronouns	-.034	-.025	.019	.041
Usage of first-person plural pronouns	.031	.001	.003	-.035
Usage of second-person pronouns	-.053	.030	-.009	.031
Usage of third-person singular pronouns	-.061	.039	.035	-.013
Usage of third-person plural pronouns	-.021	.041	-.042	.023
Usage of impersonal pronouns	.034	-.005	-.003	-.026
Usage of articles	-.003	.003	.005	-.005
Usage of prepositions	.023	.006	.036	-.065
Usage of auxiliary verbs	-.005	-.004	.030	-.021
Usage of adverbs	.015	.030	-.059	.015
Usage of conjunctions	-.031	-.013	-.001	.045
Usage of negations	-.014	-.003	.039	-.021
Usage of insight words	-.001	.014	.001	-.014
Usage of causation words	-.024	-.036	.046	.014
Usage of discrepancy words	-.011	-.013	-.012	.036
Usage of tentative words	-.030	.066	-.020	-.016
Usage of certainty words	.026	.035	-.011	-.050
Usage of differ words	-.015	-.005	.022	-.002
Was r/relationships post in 70520-dataset first activity on Reddit?	-.303	.228	.020	.055