Spontaneous Discourse in Response to Advice on Reddit

Neimhin Robinson Gunning
Trinity College Dublin
16321701
nrobinso@tcd.ie

Aadesh Milind Rasal
Trinity College Dublin
22301280
rasal@tcd.ie

Tarun Singh
Trinity College Dublin
23330140
singht1@tcd.ie

Benjamin David Vaughan
Trinity College Dublin
19333871
vaughabe@tcd.ie

Pranav Raviraj Shetty
Trinity College Dublin
23332501
shettyp@tcd.ie

April 14, 2024

Abstract

This study looks at trends in the dynamics of advice-giving and advice-receiving on the Reddit forum R/AMITHEASSHOLE, where members ask the community for moral guidance. We examine publicly accessible conversations to test for relationships between language use and resulting conversational consequences, emphasising the giving of kind counsel and eschewing hostility. We employ simple binary Bag-of-Words vectors and BERT embeddings to classify texts in terms of thankfulness and animosity, as well as leveraging the forum's textual regularity to infer moral judgements. The study looks at two research questions: 'What are the textual indicators of well-received criticism?', and 'What are the textual markers of advice-seeking posts that contribute to animosity in subsequent conversations?'. Our research attempts to identify practical advice-seeking and advice-giving techniques and help comprehend the dynamics of moral debates on the internet. We find that there is no relationship between the use of second person pronoun 'you' as the sentence subject and the expression of thankfulness in comments responding to negative advice. We also find that the frequency use of 'I'/'me' in advice-seeking self-disclosure posts has no relationship to whether the ensuining comments express animosity.

Keywords: advice-seeking, spontaneous online discourse, sentiment, affect.

1 Introduction

To give difficult-to-hear advice in a way that does not foster animosity or resentment on the part of the recipient may require tact, compassion, empathy, confidence, trust. It is a desirable capacity generally for the development of deep interpersonal relationships, but also in certain professions, such as talk-therapy and medicine, delivering difficult but authoritative advice is core to the profession's function.

In this study we analyze publicly available social media discussions to unveil patterns in the dynamics of advice-giving and advice-receiving.

The R/AMITHEASSHOLE subreddit on the social media platform Reddit, is an active forum in which users are invited to divulge personal stories containing moral dilemmas, about which other users can publicly pass judgement on individuals involved in the story. Posts (either submissions or comments) to the forum are made pseudonymously by default (unless the user intentionally identifies themself in their profile) and may be made anonymously by way of a "throw-away" account. It is also noteworthy that using the subreddit is free of charge in contrast to talk-therapy or medicine.

We interpret submissions to R/AMITHEASSHOLE as a form of moral advice-seeking. Miller (2020) has examined (through questionnaire) some of the motivations behind users who frequent thematically adjacent subreddits, R/CONFESSIONS, R/OFFMYCHEST, while Seraj et al. (2021) passively studied

the emotional impact of break-ups by scraping publicly accessible submissions and comments from Reddit. We suspect that submissions to R/AMITHEASSHOLE are often motivated a genuine emotional upheaval or interpersonal conflict, similar to the break-ups studied by Seraj et al. (2021). Therefore R/AMITHEASSHOLE may, for some users, serve as a cheap, anonymous/pseudonymous alternative to talk-therapy. This leads us to enquire whether the publicly available discussions on this forum reveal connections between language usage and emotional outcomes, and whether analysis can inform future advice-givers who wish their advice to be received with appreciation, or who wish to avoid offending the receiver.

The R/AMITHEASSHOLE community has converged on some robust patterns and schematics for how both the submission and comments on the forum are written. Very frequently, commenters will summarise their moral judgement of the participants in the story using a single acronym. A comment containing NTA (Not The Asshole) (usually) implies the author of the comment feels the author of the submission (the OP)[1] has behaved properly/ethically/morally/righteously. YTA (You're The Asshole) has the opposite meaning. ESH (Everyone Sucks Here) expresses that author of the comment finds all participants in the story to have behaved reprehensibly, and NAH (No Assholes Here) acknowledges an unpleasant situation for all participants with no particular person to blame. We leverage these patterns to automate the extraction of authors' sentiments and moral judgements at scale.

2 Literature Review

This literature review explores the connections between linguistics, artificial intelligence (AI), social media analysis and psychological research. It examines sarcasm detection, sentiment analysis, mental health assessment and ethical considerations in studies. These works show the progress and obstacles in the field while highlighting how computational methods can deepen our understanding of behavior and emotions in the digital world.

A core challenge of linguistic analysis is the task of inferring cognitive, developmental, and psychological processes or states from authors of texts. In order to accurately classify textual content, we conducted a rigorous research review surrounding popular tools within the current literature for textual classification. The Linguistic Inquiry and Word Count (LIWC) has seen widespread application within this literature for inferring various processes and states from authors of different texts. The thorough selection of the dictionaries used within the LIWC 2015 program are described in detail by Pennebaker et al. (2015); however, it is important to note that the process was largely manual, being supplemented by some automated text analysis tools. The LIWC tool operates through taking some text of arbitrary length as input, and yields an array of numeric outputs. The majority of numeric outputs by the program are normalised counts of words from various categories, each of which can indicate underlying psychological states of the author. Some of these states include positive emotions, or more nuanced states such as self-focus or anxiety. While an incredibly simple and power tool for analysing English texts at scale, the LIWC tool has one primary drawback: it disregards the *compositional* nature of language. Attempts to extract this compositional nature have been explored by Neviarouskaya et al. (2010), developing a rule-based sentiment analysis program that attempts to respect compositional semantics, along with employing a hierarchical classification of sentiment.

Another reputable tool utilised within the linguistic analysis research body is VADER (Hutto and Gilbert, 2014), a program designed to effectively analyse emotions in social media posts. VADER focuses on social media language, utilising quantitative methods to measure the intensity of emotions and the direction of sentiment. Unlike LIWC, VADER incorporates heuristics that take into account specific language nuances such as grammatical quality and syntax construction; this is particularly valuable for Reddit interactions, which are often long and complex. The dictionary employed by this tool was developed in accordance with human ratings and judgement, resulting in high correlations with human

[1] original poster

sentiment evaluations. VADER is particularly suitable for our research, being a tool tailored specifically towards social media usage.

With the prolific advent of neural networks in modern research and technology, it is without surprise that these novel and powerful tools have been applied to linguistic research. The research carried out by Finch and Chater (1992) revealed the capabilities of encoding textual content as vectors, being able to retain the original structure and semantics of text. With the increased capacity of novel neural network-based models, large collections of high-dimensional (i.e. greater than 200) vector representations of text are now capable of being processed in tractable times. An early yet notable application of this vectorisation and neural network-based approach is presented by Mikolov et al. (2013), showcasing the capabilities of the now well-established word2vec technique. Here, word2vec demonstrated that vector representations learned from unsupervised data can express semantic and structural relationships between words in intuitive algebraic relationships. Additionally, analyses on the similarity of two vector embeddings of text can also be performed through the use of techniques such as cosine similarity. A prime example of the impact of novel deep learning techniques is showcased by Garg (2023), in which an extensive study analysing the relationship between social media activity and mental well-being is conducted. Through machine learning and deep learning methods, they were able to categorise health conditions such as stress, depression, and suicidal thoughts by analysing data extracted from social media posts. This research, based on an analysis of 92 articles, identified various data extraction approaches and classification methods, all while showcasing the advancements in state of the art artificial intelligence models. An advanced neural network-based method of note is BERT, representing a significant advancement in sentiment analysis. Maas et al. (2011) showcased the integration of unsupervised learning techniques, which reflects the progression in sentiment analysis methodologies. BERT's capabilities were also made apparent in Garg's (2023) comprehensive study. Through leveraging BERT and similar techniques, researchers are capable of extracting nuanced insights from textual data, providing a method for more accurate and insightful analyses of sentiment and emotions in various contexts.

Social media research has also emerged as a rich area of study in recent years, offering valuable insights into human behaviour, interaction dynamics, and mental health issues. Miller (2020), Proferes et al. (2021), and Adelina et al. (2023) all delve into various aspects of social media research, including personal disclosure dynamics, data ethics, and support mechanisms within online communities. For example, Miller investigated the motivations behind intimate self-disclosure on platforms like Reddit. Adelina's study on support seeking and provision on Reddit is of particular relevance to our research, revealing how digital interactions shape the nature and quality of support exchanged, highlighting the role of content and tone in cultivating supportive online commmunities.

3 Research Questions

Miller (2020, p. 56) suggested investigating the motivations behind engagement with confessionary public social media postings, in particular whether it is empathetic concern or sensationalism/"shock value" that drives engagement. We seek to answer similar questions but in a slightly different forum, specifically in public spontaneous advice-seeking discussions with a confessionary bent on social media. Also, we take an interest in different types of engagement, empathetic/compassionate/productive engagement, in contrast with sensationalist/antagonistic/trolling engagement.

RQ1: Especially interesting to us are the cases where advice-seekers are given 'bad news' (in particular, a negative moral judgement of the advice-seeker's character or behaviour), and yet show appreciation towards the person who delivers the 'bad news'. We contribute to answering the question: What are the textual markers and schematics of well-received, yet negative, advice-giving comments?

RQ2: On the other side of the spectrum, we take an interest in cases where the public discussion devolves into antagonistic and hateful rhetoric. What are the textual markers and schematics of advice-

seeking posts that coincide with threads developing poorly? In other words, can we extract specific words, phrases, patterns of discourse that one should avoid when genuinely seeking compassionate and nuanced moral advice on public social media fora?

3.1 Hypotheses

3.1.1 Hypothesis 1 (RQ1)

Sinek's (2019) popular business oriented series of books contains polemic advice for business leaders, including advice about how to ask questions. Sinek (2019) recommends avoiding questions that start with 'why', e.g. prefer "What is it about that story that really matters to you?" to "Why does that story matter to you?". His theory is that the latter question "triggers the part of the brain that is not responsible for language", i.e. the emotional centres of the brain.

When we consider the difference between the two questions "Why did you do that?" and "What caused you to do that?", we expect the latter question is more likely to be received gently, causing less antagonistic emotional reaction than the former. Our reasoning, whose implication coincides with Sinek (2019), is that the former question has 'you' as the subject of the question, emphasising the agency and therefore the responsibility of that person in the events. The second question has 'what' as the subject, where 'what' is a question word whose meaning is left blank for the interrogated individual to fill in. The latter question places greater emphasis on the environment, which makes the interrogated individual feel like they are out of the spotlight and thus less antagonised. Based on this type of dynamic we hypothesise that the use of second person pronouns as the subject of sentences corresponds with greater levels of animosity in the discussion, and a lower probability of observing thankfulness in the second comment ($d = 2$).

Considering a triple of documents (c_0, c_1, c_2), where is c_0 is a submission by author a_1, c_1 is a response by auther a_2 to c_0, and c_2 is a response by a_1 to c_1, we test whether the use of the pronoun 'you' as the subject in first-level comments (c_1, TLC) is associated with a decrease in expressions of thankfulness in the corresponding second-level comments (c_2).

Null Hypothesis (H_0): The use of the pronoun 'you' as the subject of a sentence in top-level comments (c_1) by someone other than the original poster is statistically independent of whether thankfulness is expressed in the original poster's response to c_1.

Alternative Hypothesis (H_1): The use of 'you' as the subject in first-level comments (c_1) is correlated (either positively or negatively) with an expression of thankfulness in the second-level comments (second-level comment by original poster) (c_2).

3.1.2 Hypotheses 2 (RQ2)

The count of first-person pronouns in submissions (c_0) could suggest a measure of personal responsibility, self-focus, or self-awareness. This study, therefore, aims to test whether the number of first-person singular pronouns ('I' and 'me') in first comments is a predictor of follow-up social feedback, included animosity in responses to the comments and the proportion of received downvotes.

Null Hypothesis (H_0): The first-person pronoun count in initial comments (c_0) has no effect on the subsequent social feedback, specifically the count of comments with animosity and the total number of downvotes.

Alternative Hypothesis (H_1): A higher/lower count of first-person pronouns in initial comments (c_0) correlates with a decrease/increase in both the count of comments with animosity (y) and the proportion of downvotes (z), which suggests that self-focused language may have a postive impact on the social reception of the comment.

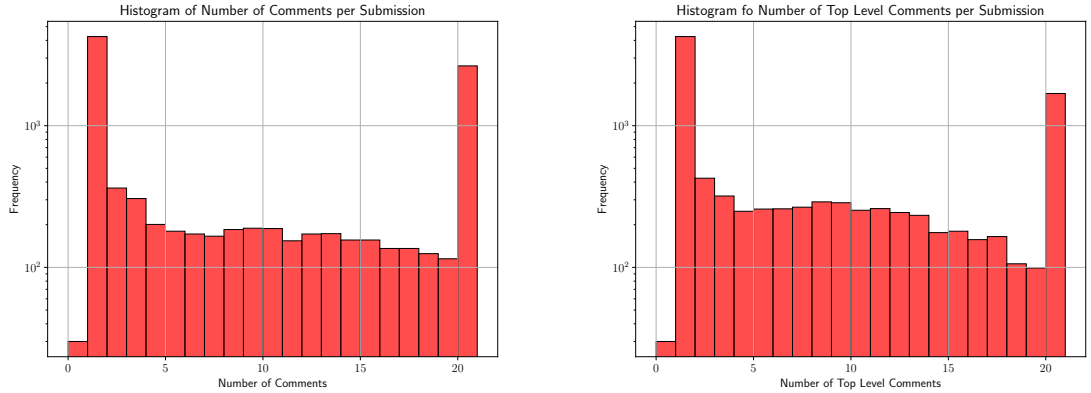


Figure 1: Distributions of number of comments on each submission, and number of TLCs on each submission. Values greater than or equal to 21 are grouped into one bar. 10,200 examples are presented. 30 examples had zero comments.

4 Research Methods

4.1 Data Retrieval

The Pushshift project is an ongoing effort to archive Reddit usage Baumgartner et al. (2020), with torrents of the entire history of Reddit available consisting more that 2 TB compressed data. Reddit User `u/Watchful1` (2023) has processed this massive dataset and made available a torrent of the 40,000 most popular subreddits, which we use to access the history of `r/AmItheAsshole` in a compressed form, without having to download the entire Reddit history. The 'info hash' for this torrent is `56aa49f9653ba545f48df2e33679f014d2829c10`. This torrent includes two files of interest to us, `AmItheAsshole_submissions.zst`, and `AmItheAsshole_comments.zst`, which compressed files with each line containing a JSON format string representing subreddit submissons and comments respectively.

The JSON format of the 'submissions' often includes a `name` field, which is referenced by the `link_id` field in the JSON representation of comments. We find 1,702,055 unique names in the set of submissions. We ignore submission records missing the `name` field, and which have been either deleted or removed (which may be a source of sampling bias). The distributions of numbers of comments and numbers of top level comments for a sample of 10,200 submissions sampled are presented in Figure 1. We find that 902 submissions have been deleted by the user, and 6,536 submissions have been removed by moderators.

We first find all submission names associated with submissions that have neither been deleted by the user (`selftext` = "$[deleted]$") nor removed by a moderator (`selftext` = "$[removed]$"), which brings us from 1,702,055 down to 468,649 submissions. A rather large number of submissions have been either deleted or removed, which is perhaps to be expected given the forum is frequently used for intimate self-disclosure. of these 468,649 submissions we uniformly sample 50,000 for further processing. Next we gather all of the comments associated with each of the 50,000 submissions, extracting them from `AmItheAsshole_comments.zst` based on the `parent_id` field matching the `name` of the submission. This results in a collection of 5,262,806 unique comments.

4.1.1 RQ1 Dataset

To answer RQ1 we need examples of advice-seeking submissions with a negative comment (namely where the comment contains either YTA or ESH). Each comment is directly descended from one other node in the discussion tree. For this dataset we are interested in comments descended directly from the first submission, i.e. the root of the tree. We say the root node (the submission) has depth 0, $d = 0$, a

comment directly on the submission has depth 1, $d = 1$, and so on. Comments with depth 1 are named TLCs (Top Level Comments) by Bao et al. (2021). We call comments with depth 2 2LCs (2nd Level Comments). Secondly, we need to see that the OP has responded to the negative comment. Therefore a sample in this dataset has three texts, the original submission c_0 authored by a_1 ($d = 0$), a comment c_1 on the original submission containing either of the substrings "YTA" or "ESH" authored by $a_2 \neq a_1$ ($d = 1$), and a comment c_2 authored by a_1 ($d = 2$) on the comment c_1 authored by a_2. We filter out triplets (c_0,c_1,c_2) in which any comment has been either deleted or removed, as before. After processing the 50,000 submissions and 5,262,806 associated comments we are left with 29,939 triplets following the above structure. Such that individual data samples are reasonably de-correlated we extract only one sample from each discussion tree (submission), leaving us with 12,649 triplets.

We're looking for differences in c_1 given c_2 expresses thankfulness, and c_1 given c_2 does not express thankfulness. Our first hypothesis is that when c_1 contains more second person pronouns it is associated with greater accusatoin, and therefore less appreciation in c_2. Our second hypothesis is that *information sharing*, which has been identified by Bao et al. (2021) as a prosocial action, will be associated with higher levels of `thankfulness` in c_2.

4.1.2 RQ2 Dataset

Out of the initial 50,000 submissions we uniformly sample 1000 submissions, resulting in a dataset containing 124,383 rows (1000 submissions + 123,383 comments). 5520 entries are of the comments had been deleted, around 4.4 percent of the data, and one entry is missing due to data corruption. After removing these we have a sum total of 118,862 rows.

Our second research question is about the process of `r/AmItheAsshole` discussions descending into negative, angry, antagonistic sentiment. For each discussion we run our automated `animosity` classification on each document. We consider the total sum of all of the `animosity` score. Additionally, we record the net amount of votes for the entire discussion. Furthermore, this exploration extends to assessing the correlation between the occurrence of first-person singular pronouns in the original submission post (c_0) and the collective measures of animosity and net votes throughout the comment threads.

4.2 Estimating `thankfulness` and `animosity`

Towards answering our research questions we focus on two qualitative extra-textual variables, `thankfulness` and `animosity`, which we attempt to estimate on a large number of documents by quantitative means. To assess the robustness of our quantitative estimations we manually label a validation data set in the following manner. 100 documents are manually contrived by the authors, and 100 documents are randomly selected from the set of submissions ($d = 0$), 102 from the set of first comments ($d = 1$), and 50 from the set of second comments ($d = 2$). Finally, because the number of sampled comments expressing thankfulness was low, we sample 52 comments containing the string 'thank'. Each document is labeled by each of the authors of this paper (5) in a binary fashion for both `thankfulness` and `animosity`, with the instruction that `thankfulness` $= 1$ implies the document expressed at least a moderate degree of thankfulness/appreciation, and similarly for `animosity`. For each document and variable this process yields five judgements (from the five authors), and our final label is the mode of these five judgements. These data are used in both calibrating and validating (using cross-fold validation) our estimation procedures.

4.3 Classification

Our classification procedure is based on two types of feature, BERT embeddings, and binary indicators of word presence (binary Bag-of-Words). We use the HuggingFace 'bert-base-uncased' model. BERT embeddings are constructed based on the special CLS (a.k.a. Beginning-of-Sentence) token emitted by the BERT tokenizer. For each document we concatenate the last four hidden states associated with the

CLS token, with each hidden state in $\mathbb{R}^{768}$, giving us a vector representation of the document in $\mathbb{R}^{3072}$. On top of these we use binary Bag-of-Words features constructed from our set of 252 manually annotated documents. We remove tokens 'nta', 'yta', 'esh', and 'nah' from the set of binary BOW features. There are 1820 tokens in the manually annotated documents resulting in binary BOW features in $0, 1^{1802}$. This gives us a vector of size $3072 + 1802$. We then apply an automated procedure for selecting a subset of meaningful colmns from this feature set. For each of `thankfulness` and `animosity` we train a logistic regression classifier using all 252 labeled samples and with 'l1' regularization parameter $C = 2$. We then select all input columns associated with a non-zero model coefficient in the logistic regression classifier. Finally, we transform the subset of selected features polynomial with degree 2.

To estimate `thankfulness` and `animosity` of a document with BERT we first tokenize the document, generate a BERT embedding for each token, and then take the mean of all of the tokens' embeddings. Using these mean embeddings as inputs to a linear model, we train on the HL1 set with aggressive 'l2' regularization (i.e. Lasso); Coates and Bollegala (2018) has argued that averaging the token embeddings of a documuent is a valid technique for the construction of meta-embeddings (i.e. fixed-length representations of the entire document). The 'l2' type of regularization causes most of the input features to be given 0 weight, which will leave us with a model that essentially selects a small number of the most important elements in the embedding. The output of the linear model is a vector of length 2: $[\texttt{thankfulness}, \texttt{animosity}]$.

Let d be the length of the vector representation of a document, i.e. $e' = [e_1, e_2, ..., e_d]$, where d is the number of features after feature reduction and polynomial transformation of the BERT and BOW features. We prepend a dummy feature to the embedding to get $e = [1, e_1, e_2, ..., e_d]$. Our linear model has $d + 1$ weights, $w_{\texttt{var}} = [w_0, w_1, ..., w_d]$, for each variable $\texttt{var} \in \{\texttt{thankfulness}, \texttt{animosity}\}$. The classification for a given variable is just:

$$s_{\texttt{var}} = \text{sign}(\sum_{i=1}^{d+1} w_{\texttt{var},i} \cdot e_i) = \text{sign}((w_{\texttt{var}})e^T)$$

This means we can interpret $w_{\texttt{var},i}$ as the importance of the associated feature e_i in discriminating the variable.

5 Results

5.1 Inter-Rater Reliability Analysis Using Krippendorff's Alpha

We employed Krippendorff's alpha as a measure of inter-rater reliability for our manually annotated supervised dataset. This data set, consisting of 252 comments, was assessed by five annotators for two attributes: Thankfulness and Animosity. Krippendorff's alpha is the most widely used statistical measure of the level of agreement between multiple raters and adjusts for the amount of agreement that would be expected merely by chance.

We observed high agreement in annotation analysis for both attributes. For Thankfulness, the obtained Krippendorff's alpha value was 0.8909. The high value of alpha refers to strong consistency among the annotator's ratings. Whereas for Animosity, the calculated coefficient was 0.7762, which indicates substantial agreement between the raters.

5.2 Classifier Tuning and Evaluation

To train our classifiers of `thankfulness` and `animosity`, we first apply an automated method for selecting a subset of the BERT and binary Bag-of-Words features. We use the entire manually annotated set to train a logistic regression for each variable, with 'l1' regularization parameter $C = 2$. The trained models' coefficient vectors are sparse. We take only the features with non-zero coefficients. For the thankfulness classifier this results in selecting 45 of the BERT embedding columns (out of 3072), and 13 features representing the presence of words (binary Bag-of-Words). For the animosity classifier

	feature name	logistic regression coefficient
0	and	0.2692876333573884
1	another	2.871375438394005
2	can	-0.34322029469575155
3	comments	1.8041973359019075
4	do	-0.5257069949394588
5	don	-0.017187268022520807
6	get	2.203756341833472
7	in	-0.445682394419448
8	just	0.05398864339687095
9	like	-0.3742718549158877
10	look	0.05127208637082296
11	my	-0.117507347809231
12	now	0.3806345235897748
13	on	-0.35259721225921
14	really	-0.4537797754832705
15	said	0.060405757348280936
16	since	0.5505404692440539
17	time	0.14296918448294066
18	very	-0.24473184438593334

	feature name	logistic regression coefficient
0	always	0.6465228020541558
1	appreciate	0.21273699486220415
2	back	0.0971532445167887
3	be	-1.3663897665014506
4	for	0.7884249564855321
5	it	0.13714428946130605
6	tbh	0.7696372753836772
7	thank	3.698054144369798
8	thanks	1.0542837946334294
9	think	1.942552513565848
10	to	0.3449108605736806

Table 1: Automatically selected words for binary BOW features. Left animosity, right thankfulness.

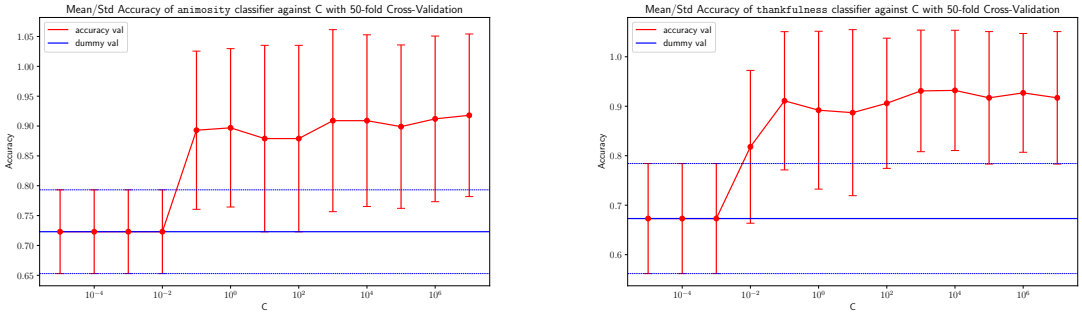


Figure 2: Tuning 'l1' regularization term C for logistic regression classifiers of `animosity` and `thankfulness`

we end up selecting 69 BERT embedding columns and 14 BOW columns. The words selected for inclusion in the BOW features for each classifier are presented in Table 1. For the animosity classifier the most discriminating words are 'another', 'comments', and 'get'. The words chosen in this automated procedure for the animosity classifier seem strange and unintuitive. For thankfulness classifier the most discriminating are 'thank' and 'thanks', 'think', with 'think' being oddly more discriminating than 'thanks'.

The selected features ($59+11 = 70$ for thankfulness, $68+19 = 87$ for animosity) are then transformed polynomially with degree 2, i.e. for pair of features x_i, x_j we create new features: $[x_i x_j, x_i^2, x_j^2]$, resulting in feature vectors of length $\frac{(n+d)!}{n!d!}$ (2555 for thankfulness, 3915 for animosity) including dummy feature for the intercept.

With these feature sets we then perform 50-fold cross-validation on the training dataset (80% of manual annotations) for each classifier, with varying 'l1' regularization C. This means that for each tested hyperparameter C we train 50 logistic regression models on different subsets of the training data, keeping a portion for validation. For each parameter we present the mean accuracy on the 50 validation sets in the plots in Figure 2. The accuracy is compared to the accuracy of a baseline classifier that always predicts the most frequent class in the training data. The mean accuracy of the baseline classifier is plotted as a continuous blue line in Figure 2, with standard deviation indicated by dotted blue lines.

Based on these cross-validation results we select $C = 10^4$ for both classifiers. We then train each classifier again on the full 80% training set, and present summary evaluation metrics of the two models computed on the 20% test data in Table 2. For the animosity classifier there is an improvement of about 15.5% accuracy compared to the 'most-frequent' baseline, whereas for the thankfulness classifier it is an improvement of about 25.5%. Given the levels of disagreement between annotators we consider the accuracy of the classifiers acceptable for our purposes.

The animosity classifier is balanced in terms of precision and recall, whereas the thankfulness classifier

	metric	animosity most frequent	animosity logreg	thankfulness most frequent	thankfulness logreg
0	accuracy	0.745	0.836	0.691	0.945
1	precision	0.0	0.667	0.0	0.889
2	recall	0.0	0.714	0.0	0.941
3	f1	0.0	0.69	0.0	0.914

Table 2: Evaluation scores of the two models, compared to a baseline model that always predicts the most frequent class, evaluated on the held out 20% of the manually annotated data.

	doc	logreg thankfulness	manual thankfulness
0	Okay thanks, I will look into that.	0	1
1	Every time I think it can't get worse, you prove me wrong.	1	0
2	Your moral counsel is genuinely approcieated	1	1

Table 3: Example documents along with manual and model classifications of thankfulness.

has better recall than precision. The recall metric is $tp/(tp + fn)$, whereas precision is $tp/(tp + fp)$, so the thankfulness model is more likely to make a false positive error than a true positive error.

We present some example classifications of the thankfulness model in Figure 3. The first sentence in this table contains the word "thanks" and should have been easy for the model to classify, but it is misclassified. The second sentence is another example of a misclassification by the model, possibly occurring due to the phrase "you prove me wrong" being typically positive. The final example contains a misspelling "approcieated" and yet is correctly classified by the model.

The models used later in this work are trained on the full set of manually annotated labels with the selected hyperparameter $C = 10^4$, in order to maximize supervision of the classifier. However, this means we do not present an evaluation of the particular model used in the later analyses.

5.3 RQ1 Results

In the set of 50,000 submissions selected at random, we found 12,649 'threads' matching our criteria for inclusion. Let t represent the second comment (2LC) being classified as thankful, and y represents the first comment (TLC) containing the second person pronoun 'you' as subject of a sentence. The contintency table for counts of these events and their negatives are presented in Table 4. We test the null hypothesis H_0 that counts of t and y are independent of each other and test with significance level $\alpha = 0.05$. Using Pearson's χ^2 test of independence we observe $\chi^2 = 0.12284$ and $p = 0.726$, indicating that their is no evidence to reject the null hypothesis.

5.4 RQ2 Results

The statistical analysis from the research, utilizing Spearman's rank correlation coefficient, indicates that the frequency of "I" and "me," particularly in the initial comments, has no statistical effect on the chances of being downvoted or receiving animosity. Specifically, the calculated statistic for downvotes was 0.0047, with a p-value of 0.882, and for 'I/Me' and animosity was -0.0175, with a p-value of 0.580. With such results, they would reject the alternative hypothesis, which states that a higher or lower count of first-person pronouns relates to difference in social feedback.

	$\neg t$	t
$\neg y$	9226	2267
y	923	233

Table 4: Contingency table of events t and y for the RQ1 dataset.

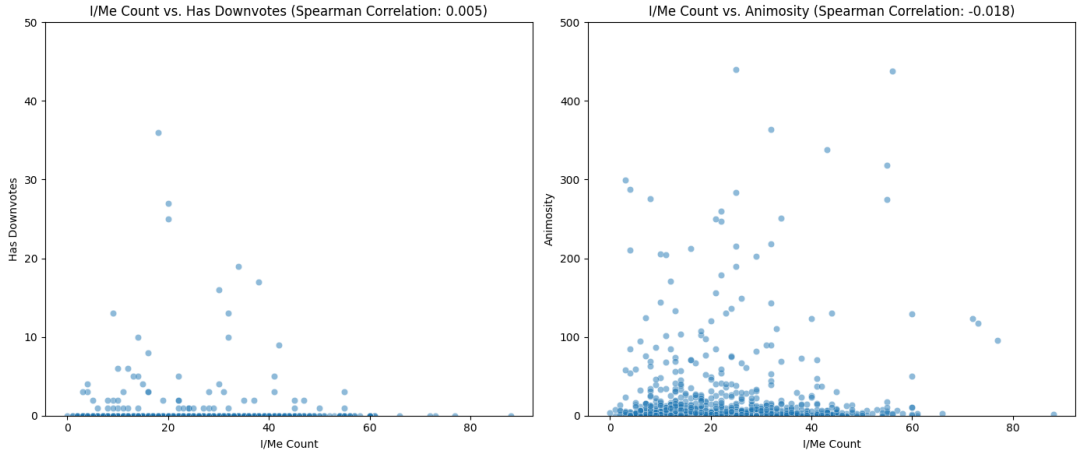


Figure 3: Scatterplot of I/Me Count vs Downvote (left) and I/Me Count vs Animosity (Right)

Therefore, this study supports the fact that from the data, there is no substantial evidence that the first person, by itself, would play any role in how the comment is received in terms of down-votes and animosity. These findings highlight that the remaining elements of the comments or contextual elements may have a greater influence on social feedback. Thus, from the null hypothesis, the hypothesis does not predict that the amount of social feedback that ensues from the quantity of first-person pronouns used.

6 Conclusions

Our study investigated the dynamics of advice-giving and advice-receiving on the R/AMITHEASSHOLE subreddit. We find that there is no correlation between the presence of 'you' as the subject of a sentence in the first comment (TLC) and the thankfulness expressed by the subsequent comment (2LC). In terms of extracting practical advice for advice-givers, this means that our study suggests that addressing the advice-seeker directly with 'you' and emphasing the agency of the advice-seeker by using 'you' as the subject of sentences will not, in itself, influence the advice-seeker's reception of the advice.

Similarly, we also find that the use of first person singular pronouns 'I' and 'me' in advice-seeking posts has no correlation with our proxy measures of the overall animosity of the ensuing discussion. In terms of practical advice, our study suggests that advice-seekers can use first person singular pronouns liberally and it will not, in itself, trigger greater animosity by advice-givers and commenters.

The R/AMITHEASSHOLE subreddit is a rich source of information about conversational, moral, and social dynamics, of which we have only scratched the surface. The textual regularity self-imposed by the community allows for simple and robust inference of the users' beliefs and moral judgements, as well as some demographic info, which provides an opportunity for correlation with other aspects of the text.

References

Adelina, N., C. S. Chan, K. Takano, P. H. M. Yu, P. H. T. Wong, and T. J. Barry (2023). The stories we tell influence the support we receive: examining the reception of support-seeking messages on reddit. *Cyberpsychology, Behavior, and Social Networking 26*(11), 823–834.

Alfandre, D. J. (2009). "I'm going home": Discharges against medical advice. *Mayo Clinic Proceedings 84*(3), 255–260.

Bao, J., J. Wu, Y. Zhang, E. Chandrasekharan, and D. Jurgens (2021). Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*, pp. 1134–1145.

Baumgartner, J., S. Zannettou, B. Keegan, M. Squire, and J. Blackburn (2020). The pushshift reddit dataset. *CoRR abs/2001.08435*, 1–11.

Botzer, N., S. Gu, and T. Weninger (2021). Analysis of moral judgement on reddit.

Coates, J. and D. Bollegala (2018). Frustratingly easy meta-embedding–computing meta-embeddings by averaging source word embeddings. *arXiv preprint arXiv:1804.05262*, 1–5.

Finch, S. and N. Chater (1992). Bootstrapping syntactic categories using statistical methods. In *14th Annual Conference of the Cognitive Science Society*, pp. 229–235.

Garg, M. (2023). Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering 30*(3), 1819–1842.

Hutto, C. and E. Gilbert (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media 8*(1), 216–225.

Joshi, A., P. Bhattacharyya, and M. J. Carman (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR) 50*(5), 1–22.

Maas, A., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*, 1–12.

Miller, B. (2020). Investigating reddit self-disclosure and confessions in relation to connectedness, social support, and life satisfaction. *The Journal of Social Media in Society 9*(1), 39–62.

Neviarouskaya, A., H. Prendinger, and M. Ishizuka (2010). Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pp. 806–814.

Pennebaker, J. W., R. L. Boyd, K. Jordan, and K. Blackburn (2015). The development and psychometric properties of liwc2015. Technical report, The University of Texas Austin.

Proferes, N., N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society 7*(2), 1–14.

Reddit User `u/Watchful1` (2023). Separate dump files for the top 40k subreddits. Available on `academictorrents.com`[2]. Originally posted on Reddit[3]. Last accessed 28th Feb '24.

Seraj, S., K. G. Blackburn, and J. W. Pennebaker (2021). Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proceedings of the National Academy of Sciences 118*(7), 1–7.

Sinek, S. (2019). *Find your why*. Gramedia Pustaka Utama.

[2] https://academictorrents.com/details/56aa49f9653ba545f48df2e33679f014d2829c10

[3] https://www.reddit.com/r/pushshift/comments/1akrhg3/separate_dump_files_for_the_top_40k_subreddits/

7 Group Member Contributions

7.1 Neimhin Robinson Gunning, 16321701

Neimhin contributed summaries for several papers referenced in this work, which were then incorporated into the literature review. After Pranav contributed a first draft of the Literature Review, Neimhin contributed significant edits, additions, and restructurings.

Neimhin wrote the first draft of the Introduction section.

Neimhin wrote the first draft of the Research Questions section, and the work-in-progress high level hypotheses at the end of that section.

Neimhin contributed details and specificity to the Research Methods section, including how the raw dataset was obtained. Neimhin wrote a first draft of the Analysis subsection of the Research Methods subsection. Neimhin wrote the first drafts of the RQ1 Dataset and RQ2 Dataset subsubsections.

Neimhin maintained meta data relating to literature reviewed for this study. The metadata are provided in the associated replicability archive in `./additional-data`.

Neimhin wrote the following source code files for data preparation, model fitting, and analysis (included in the replicability archive):

```
./src/bert_cls.py
./src/can_use.py
./src/cls.py
./src/extract_comments.py
./src/extract_submissions.py
./src/features.py
./src/man_embeddings.py
./src/manual_eval.py
./src/manual_train_cls_animosity.py
./src/manual_train_cls_animosity_vis.py
./src/manual_train_cls_thankfulness.py
./src/manual_train_cls_thankfulness_vis.py
./src/mk_ds1_group.py
./src/mk_ds1_pre.py
./src/mk_ds1_sample.py
./src/mk_ds1_valid.py
./src/ds1-cls-only.py
./src/ds1.py
./src/ds1-test.r
./src/ds1_embeddings_c2.py
./src/polynomial.py
./src/select-50000-submissions.py
```

7.2 Benjamin Vaughan, 19333871

Ben contributed summaries for 2 papers referenced in this work, namely Adelina et al. (2023); Proferes et al. (2021). Ben supported Aadesh during his construction of the methodology section, discovering a paper highly related to our research question (Botzer et al., 2021), and additional techniques that could be used in classification tasks (Hutto and Gilbert, 2014). Ben reconstructed and re-synthesised the literature review section, augmenting the description of our readings in order to better support our hypotheses. Ben was also involved in the exploration of RQ2, aiding in question definition and definition, along with providing support regarding the construction of the RQ2 dataset.

7.3 Tarun Singh, 23330140

Tarun's involvement in this project included summarizing three referenced papers; Maas et al. (2011), Alfandre (2009) and Singh (2023) and integrating these summaries into the literature review section. Tarun also helped to revise the literature review and collaborated with Pranav on the review. Furthermore, Tarun assumed responsibility for writing the paper's abstract. Tarun helped with file system administration,

making sure that scripts and data were easily accessible and arranged. Tarun was essential in helping to handle data and test hypotheses while writing scripts for RQ2 analysis.

Tarun helped Neimhin implement the stanza classifier and helped construct a reliable pipeline for preprocessing, cleaning, and data feeding.

7.4 Pranav Raviraj Shetty, 23332501

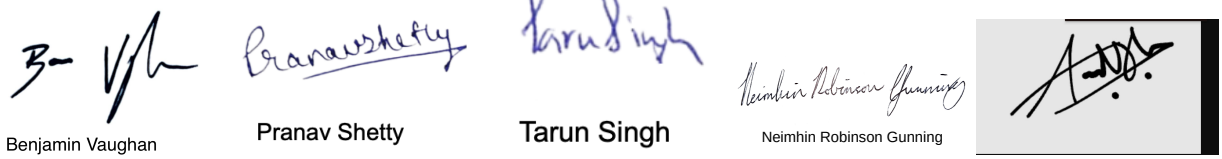
Pranav contributed summaries for 3 papers referenced in this work (Joshi et al., 2017; Garg, 2023; Hutto and Gilbert, 2014). Pranav prepared the first draft of the literature review and then collaborated with Neimhin in restructuring it. Pranav created the RQ2 dataset for analysis and carried out the pearson correlation test on the RQ2 dataset to get the final results. Pranav also wrote the hypothesis (first-draft) and RQ2 results section of the paper. Pranav made the following files:

```
creating_1000submissions_with_comments.py
splittingthedataset.py
ds2_embeddings_1.py
ds2_embeddings_2.py
ds2_animosty.py 1
mergingfile
RQ2-MeCount.py
Merger.py
finalmergerRQ2.py
spearman.py
```

7.5 Aadesh Milind Rasal, 22301280

Aadesh contributed summaries of 2 papers related to the original topic of the effect of medical advice on individuals. Aadesh prepared research methods which include **Data Retrieval**, **Classification**, and **Evaluation**. Aadesh helped in producing the initial analysis of RQ1 dataset by applying preprocessing techniques (stopwords, stemming, pruning text to keywords) deduced the top 20 most frequent words and with manual analysis deduced the thankful and animos comments from the 200 sample comments , evaluated random forest and support vector classifiers, assisted in manual annotations of the dataset

7.6 Signatures



Benjamin Vaughan | Pranav Shetty | Tarun Singh | Neimhin Robinson Gunning