Classifying Reasonability in Retellings of Personal Events Shared on Social Media: A Preliminary Case Study with /r/AmITheAsshole

Ethan Haworth[1], **Ted Grover**[2], **Justin Langston**[1], **Ankush Patel**[1], **Joseph West**[1], **Alex C. Williams**[1]
[1]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville
[2]Department of Informatics, University of California, Irvine
ehaworth@vols.utk.edu; grovere@uci.edu; jlangst6@vols.utk.edu;
apatel79@vols.utk.edu; jwest60@vols.utk.edu; acw@utk.edu

Abstract

People regularly share retellings of their personal events through social media websites to elicit feedback about the reasonability of their actions in the event's context. In this paper, we explore how learning approaches can be used toward the goal of classifying reasonability in personal retellings of events shared on social media. We collect 13,748 community-labeled posts from /r/AmITheAsshole, a subreddit in which Reddit users share retellings of personal events which are voted upon by community members. We build and evaluate a total of 21 machine learning models across seven types of models and three distinct feature sets. We find that our best-performing model can predict the reasonability of a post with an F1 score of .76. Our findings suggest that features derived from the post and author metadata were more predictive than simple linguistic features like the post sentiment and types of words used. We conclude with a discussion on the implications of our findings as they relate to sharing retellings of personal events on social media and beyond.

Introduction

People regularly share retellings of their personal events through social media websites. The primary motivation to share these retellings on social media is, by and large, centered around receiving feedback about the *reasonability* of one's own actions within the purview of a personal event or experience, particularly when it involves other people (Ellison, Steinfield, and Lampe 2007). Assessing the reasonability of one's own actions on social media can take place in direct messages with close friends or in large forums in which anonymity is afforded. Regardless of how reasonability is assessed, social media has maintained its utility as a tool for eliciting feedback about the ethics, morality, and appropriateness of a person's actions within a given context or event (Teitelbaum and others 2020).

In this paper, we present findings from a preliminary case study aimed at understanding how learning approaches can be used toward the goal of classifying reasonability in personal retellings of events shared on social media. Our study is grounded in data collected from /r/AmItheAsshole/, a popular subreddit community with 2.5 million subscribers that allows Reddit users to post retellings of their personal events (see Figure 1) in which other Reddit users respond with comments about whether the individual acted reasonably (i.e. Not the Asshole) or unreasonably (i.e., You're the Asshole) based on the written retelling of their event. The subreddit includes a comprehensive FAQ that details both how community members should write new submissions and how they should go about voting on posts. Recent research has used the subreddit's data to reliably examine various facets of human judgements, reinforcing its reliability as a tool for research (Hu, Whiting, and Bernstein 2021).

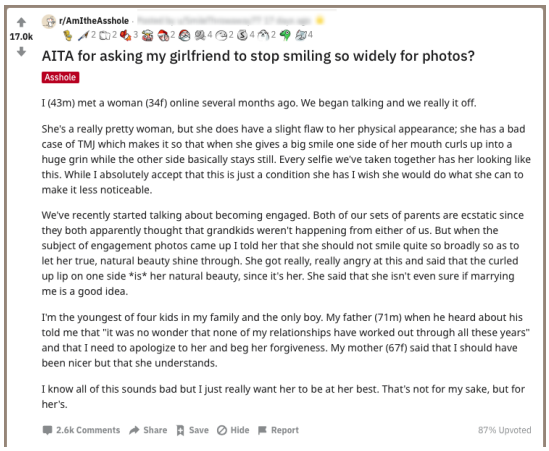

Figure 1: An example post from /r/AmITheAsshole.

Using a dataset of 13,748 subreddit posts labeled by the subreddit community as YTA or NTA, we build, tune, and evaluate a total of 21 binary classification models, exploring seven model types across three sets of features. We find that our best-performing model – a Random Forest classifier – achieves an F1 score of 0.76. Our findings suggest that model-based approaches can achieve high accuracy when trained on datasets that include metadata about the post and its author, rather than linguistic features of the post itself. We conclude by discussing the practical implications of modelling retellings of personal events on social media sites.

Related Work

Our work explores how learning methods can be used to classify whether an individual acted reasonably based on a retelling of a personal event posted on social media. We describe the prior work related to sharing personal events on social media and approaches for modelling reasonability.

Sharing Personal Events on Social Media

People regularly share retellings of their personal events through social media websites. The primary motivation to share these retellings on social media is, by and large, centered around receiving feedback about the *reasonability* of one's own actions within the purview of a personal event or experience, particularly when it involves other people (Ellison, Steinfield, and Lampe 2007). Studies have shown that self-disclosing these events on social media can help individuals feel both approved and validated in their decision-making (Bazarova and Choi 2014). More broadly, the role of sharing such personal information has been shown to be substantial, providing social support that positively influences individuals' mental health and wellness (De Choudhury and De 2014; Moreno et al. 2011; Park, McDonald, and Cha 2013). Despite a wealth of research suggesting the benefits of sharing information about personal events on social media, the act of doing so remains challenging and controversial for some, particularly when their anonymity is at risk or when feedback expectations cannot be adequately assessed (Ammari, Schoenebeck, and Romero 2019).

Modelling Reasonability on Social Media

Research has devoted significant attention to modelling logical correctness in online argumentation and deliberation in social media settings (Halpern and Gibbs 2013). Significant attention has been specifically given to understanding how modeling can make predictions about the arguments that take place on the Internet and the involved persons. For example, research has investigated how modeling text and language can successfully identify strengths in convincing arguments (Habernal and Gurevych 2016), surface weaknesses in unconvincing arguments (Persing and Ng 2017), and model linguistic differences in polarized viewpoints (Trabelsi and Zaïane 2019). More fundamental examinations have made inquiries regarding how you can model persuasion and mine arguments from social media. (Dusmanu, Cabrio, and Villata 2017; Dutta, Das, and Chakraborty 2020; Misra et al. 2017). Recent contributions have introduced argumentation datasets for further exploring new computational approaches (Bosc, Cabrio, and Villata 2016; Gretz et al. 2020). The societal ramifications of these datasets and approaches is significant as machine learning models are becoming increasingly pervasive, particularly in legal settings (Aletras et al. 2016; Cormack and Grossman 2014).

In this paper, we build on this prior literature with models that classify the "*reasonability*" of peoples' actions based on retellings of their events on social media. We ground our study in a dataset collected from Reddit and present findings that suggest that machine learning models can accurately classify whether an person's actions were reasonable based on a personal event as retold by the person on Reddit.

Method

The goal of our research is to explore how learning approaches can be used toward the goal of classifying reasonability in retellings of personal events shared on social media. Here, we describe the dataset collected for our study and the modeling approach we use to fuel our research inquiry.

Data Collection Procedure

We generated a dataset of personal event retellings and their associated labels of reasonability by using the Python Reddit API Wrapper (PRAW) to scrape and extract posts from the popular subreddit /r/AmITheAsshole. By default, each post in the subreddit includes a title, a body, an author, and a series of comments from other Reddit users. Some of these comments include one of five votes to evaluate the reasonability of the original poster's actions according to their retelling of the event: (1) You're The Asshole (YTA), (2) Not The Asshole (NTA), (3) Everybody's The Asshole (ETA), (4) Nobody's the Asshole (NAH), and (5) Not Enough Info (INFO). In order to simplify our inquiry, we focus on classifying posts with the subreddit's two primary labels: YTA and NTA. This makes the space suitable for binary classification methods.

Through PRAW, we scraped a total of 211,742 posts from /r/AmITheAsshole from January 1, 2020 to June 30, 2020. We chose this time range so that we could safely assume that the voting for all extracted posts had, in fact, concluded. 113,747 (53.7%) of the extracted posts had deleted text bodies and 37,939 (17.9%) of the extracted posts had text bodies that were removed (e.g., by moderators). We assigned each of the remaining 60,056 posts a target label by applying the majority vote of the labels presented in the post's associated comments. We further supplemented the majority vote label by calculating the YTA-ratio, i.e. the number of YTA votes divided by the sum of the numbers of YTA and NTA votes.

Filtering and Sampling Procedure. We applied a two-step filtering procedure to the dataset of 60,056 posts to ensure our data included posts that had clear outcomes for YTA and NTA. First, we removed any post with no YTA or NTA labels in any of the post comments. Second, we removed any post that had a YTA-ratio that fell between the range of 0.3 to 0.7. We view posts in this range as too ambiguous for binary classification. After completing the filtering procedure, the remaining dataset included 6,874 labeled instances of YTA and 46,906 instances of NTA. To counter the observed imbalance, we randomly sampled 6,874 NTA instances to match the YTA sample size. The final dataset therefore included 13,748 posts equally balanced between the YTA and NTA classes.

Feature Sets

Using the user data and post data returned from PRAW, we designed two distinct feature sets based on the context from which they were drawn. Alongside these feature sets, we also explore a third feature set, which we refer to simply as Feature Set #3: `All`, that includes all 117 features from both Feature Set #1 and Feature Set #2.

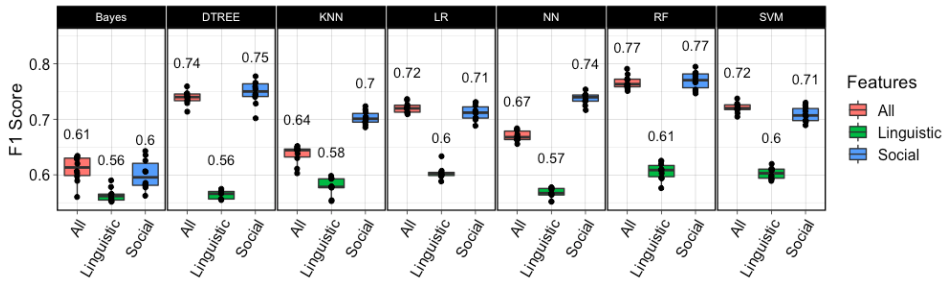

Figure 2: F1 scores for cross-validation runs of each model type trained on each feature set, annotated with the mean F1-score.

Feature Name	Feature Description
Linguistic Features	
title_uppercase_count	Num. of capitalizations in title
title_word_count	Num. of words in ttitle
title_profanity_count	Num. of profane words in title
avg_word_length	Avg length of words in post
stop_word_count	Num. of stopwords in post
numerics_count	Num. of numbers in post
uppercase_words_count	Num. of capitalizations in post
sentence_count	Num. of sentences in the post
avg_sentence_length	Avg num. of words per sentences
profanity_count	Num. of instances of profanity
Social Features	
post_score	Score of upvotes to downvotes
post_upvote_ratio	Ratio of upvotes to total votes
post_num_comments	Number of post comments
post_creation_date	Timestamp of post creation
post_over_18	Is the post flagged as NSFW?
post_edited	Has the post been edited?
author_comment_karma	Aggregated comment karma
author_link_karma	Aggregated link karma
author_has_verified_email	Author has verified e-mail?
author_acct_creation_date	Timestamp of account creation
author_is_mod	Author is a Moderator?
author_is_gold	Author has Reddit Premium?

Table 1: Social and linguistic feature names and descriptions that supplement the LIWC and SentimentR features.

Feature Set #1: Linguistic Features The first feature set includes 105 features that describe the post's linguistic nature. Most (93) of these features come from LIWC (Linguistic Word County and Inquiry) (Pennebaker et al. 2015). LIWC calculates a rate, or percentage, of words in a sample of text belonging to several predefined categories and has been reliably used in prior empirical studies on the detection of meaning in text in a wide variety of contexts. We then use the SentimentR (Naldi 2019; Rinker 2017), a powerful sentiment engine that takes into account the presence of negators in sentences to generate two features that quantify both the post title and the post body text sentiment. We complement these 95 features with 10 additional features based on descriptive textual features about each post's title and body. Table 1 details the 10 linguistic features that complement the LIWC and SentimentR features.

Feature Set #2: Social Features The second feature set includes a set of 12 features, shown in Table 1, that describe the post's social attributes extracted directly from PRAW.

Binary Classification Models

We explored seven different classification methods for evaluating this dataset: decision trees (DTREE), random forest classifiers (RF), logistic regression (LR), neural networks (NN), support vector machines (SVM), k-nearest neighbors (KNN), and naive Bayes classifiers (Bayes). We iteratively hypertuned all possible parameters for each model type using both a coarse grid search and fine grid search using Scikit-Learn. After identifying the best-performing parameters for each model type, we employed a standard 10-fold cross validation and fit the model on a 80% train and 20% test split. The models were then evaluated based on their accuracy, precision, recall, and F1 score on the heldout test set.

Results

Figure 2 shows the F1-score for all cross-validation runs on each of feature sets used for each model type. We find that RF classifiers generally out-performed the other model types across all three datasets during cross-validation. DTREE classifiers were the only model type to consistently achieve a set of competitive F1-scores to the RF classifiers. In comparison to tree-based classifiers, LR and SVM classifiers performed marginally worse while other model types performed more variably. Alongside the observations made during cross-validation, we observe similar trends in applying the best-performing model of each model type to the test set, suggesting that overfitting was a non-issue. We found that RF classifiers continued to serve as the best-performing model type across all feature sets as shown in Table 2.

We observe significant distinctions in the performance of each feature set. Models that utilized only the Linguistic feature set consistently under-performed in comparison to mod-

Feature Set	Model	Acc.	P	R	F1
Linguistic	RF	0.61	0.60	0.62	0.61
Social	RF	0.78	0.83	0.71	0.76
All	RF	0.77	0.81	0.71	0.76

Table 2: Accuracy (Acc), Precision (P), Recall (R), and F1-Score for best-performing models for each feature set.

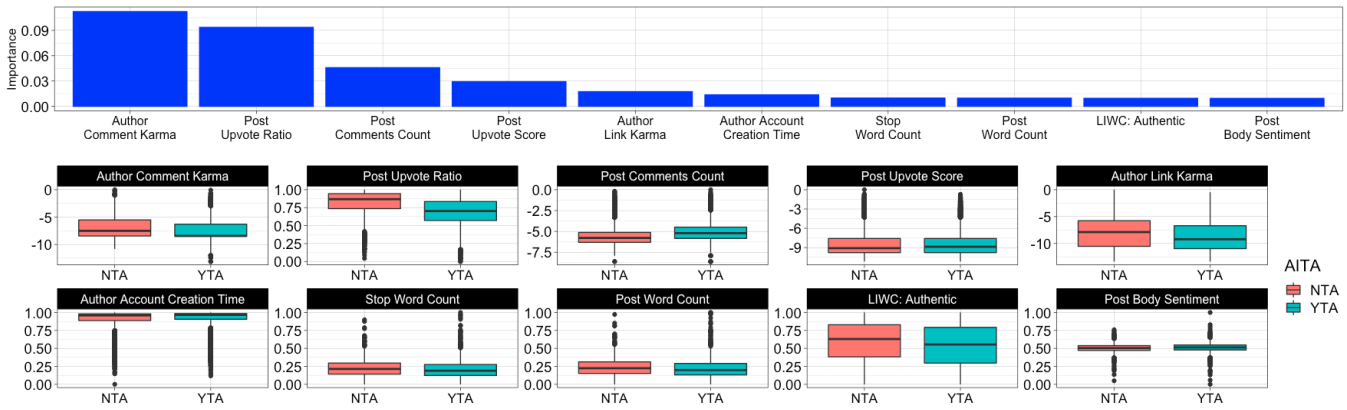

Figure 3: Importance (top) and normalized values (bottom) of the top-10 most important features. For some features above an additional log transformation was applied for ease of visualization. **** $p < .0001$ (Wilcoxon Rank Sum Test).

els that utilized either of the other datasets. This is further supported by the notion that the best-performing models' most important features stemmed from social characteristics of the subreddit posts rather characteristics of the text itself according to the model coefficients calculated by sklearn. As shown in Figure 3, we observe that posts labeled as NTA tend to have higher author comment karma, higher post-upvote ratio and score, as well as fewer comments. However, a smaller number of linguistic features were shown to be important. Posts labeled as NTA tend to use more stop words (e.g., articles), used more words overall, had a higher LIWC authenticity score, and a lower body text sentiment score.

Discussion

In this paper, we explored how reasonability can be classified in retellings of personal events on social media through community-generated and community-labeled data collected from the popular subreddit /r/AmItheAsshole/. The implications of our findings introduce new questions about the practical and ethical utility of such models as tools that act as tools for automating behavioral validation both within and beyond social media contexts.

Our research finds that social features, such as post upvote ratio, were very important in our models. We find this to be somewhat surprising, as /r/AmItheAsshole/ maintains strict rules about posts being downvoted on the basis of being labeled as NTA or YTA, and that *"downvotes should only be reserved for off-topic discussions and spam"* [1]. This finding suggests that subscribers may maintain a cognitive bias toward down-voting posts deemed YTA. Furthermore, our finding that author comment karma was the most important feature suggests that judgements of author credibility and past behavior on the platform may be more persuasive than the content of the post itself. This in part is supported by the human tendency to judge a claim by its origin rather than the claim itself (i.e. 'the fallacy of origin') (Scalambrino 2018). Alternatively, it is possible that posters with lower author karma were just more prone to provide poorer self-presentation.

An important consideration for our findings is that biases affect how personal events are both remembered and retold (Tversky and Marsh 2000). By design, /r/AmITheAsshole caters to peoples' biases in seeking out validation regardless of its necessity. As one user in the community states, *"You are so obviously NTA that I'm confused why you are asking us"*[2]. Unsurprisingly, our data collection procedure revealed that the number of NTA posts greatly outweigh the number of YTA posts. Further study is needed to better understand the nature of people engaging with social outlets, such as /r/AmITheAsshole, that vote, as a community, on the reasonability of a person's actions in a given setting and the role that anonymity plays within them (Ammari, Schoenebeck, and Romero 2019).

Our work is a preliminary case study and has limitations in terms of the data and feature set. Future work should explore building more sophisticated linguistics features (e.g. through topic modeling), design models that move beyond binary classification to classify edge-cases (e.g. Everybody's the Asshole), and conduct qualitative studies to better understand the motivations and rationales that people maintain in engaging with online communities that evaluate the reasonability of one's own actions.

Conclusion

People regularly share personal information and narratives through social media websites. In this paper, we explored how learning approaches can be used toward the goal of classifying reasonability in personal retellings of events shared on social media. Using a dataset of 13,748 community-labeled posts from /r/AmITheAsshole, We found that our models can classify the reasonability of a post with an F1 score of 0.76. We concluded with a discussion on the practical and ethical implications of our work alongside future research directions.

[1] Rule 2; https://www.reddit.com/r/AmItheAsshole/

[2] https://www.reddit.com/r/AmItheAsshole/comments/jwbono/aita_for_getting_a_former_employee_banned/

References

Aletras, N.; Tsarapatsanis, D.; Preoţiuc-Pietro, D.; and Lampos, V. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science* 2:e93.

Ammari, T.; Schoenebeck, S.; and Romero, D. 2019. Self-declared throwaway accounts on reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–30.

Bazarova, N. N., and Choi, Y. H. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication* 64(4):635–657.

Bosc, T.; Cabrio, E.; and Villata, S. 2016. Dart: A dataset of arguments and their relations on twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1258–1263.

Cormack, G. V., and Grossman, M. R. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 153–162.

De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Dusmanu, M.; Cabrio, E.; and Villata, S. 2017. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2317–2322.

Dutta, S.; Das, D.; and Chakraborty, T. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management* 57(2):102085.

Ellison, N. B.; Steinfield, C.; and Lampe, C. 2007. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of computer-mediated communication* 12(4):1143–1168.

Gretz, S.; Friedman, R.; Cohen-Karlik, E.; Toledo, A.; Lahav, D.; Aharonov, R.; and Slonim, N. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7805–7813.

Habernal, I., and Gurevych, I. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1214–1223.

Halpern, D., and Gibbs, J. 2013. Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. *Computers in Human Behavior* 29(3):1159–1168.

Hu, X. E.; Whiting, M. E.; and Bernstein, M. S. 2021. Can online juries make consistent, repeatable decisions? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Misra, A.; Oraby, S.; Tandon, S.; Ts, S.; Anand, P.; and Walker, M. 2017. Summarizing dialogic arguments from social media. *arXiv preprint arXiv:1711.00092*.

Moreno, M. A.; Jelenchick, L. A.; Egan, K. G.; Cox, E.; Young, H.; Gannon, K. E.; and Becker, T. 2011. Feeling bad on facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety* 28(6):447–455.

Naldi, M. 2019. A review of sentiment computation methods with r packages. *arXiv preprint arXiv:1901.08319*.

Park, M.; McDonald, D.; and Cha, M. 2013. Perception differences between the depressed and non-depressed users in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of liwc2015. Technical report.

Persing, I., and Ng, V. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *IJCAI*, 4082–4088.

Rinker, T. 2017. Package 'sentimentr'. *Retrieved* 8:31.

Scalambrino, F. 2018. Genetic fallacy. *Bad Arguments: 100 of the Most Important Fallacies in Western Philosophy* 160–162.

Teitelbaum, A., et al. 2020. *The Ethics of Reddit and an Artificial Moral Compass.* Ph.D. Dissertation, New York, NY. Stern College for Women. Yeshiva University.

Trabelsi, A., and Zaïane, O. R. 2019. Phaitv: A phrase author interaction topic viewpoint model for the summarization of reasons expressed by polarized stances. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 482–492.

Tversky, B., and Marsh, E. J. 2000. Biased retellings of events yield biased memories. *Cognitive psychology* 40(1):1–38.