The Journal of Social Media in Society
Spring 2023, Vol. 12, No. 1, Page 61-86
thejsms.org

Do First Encounters Make or Break New Users? Using Text Features in the First Comment to Predict New User Return on Reddit

Emma M. Klugman
Harvard University, Cambridge, MA
emma.m.klugman@gmail.com

Many new users quit a site after only one interaction. Existing studies of user return consider user characteristics and simple feedback like upvotes, while leaving potentially useful text data unstudied. Here, we analyze 700,000 first post/sole comment pairs on Reddit, with the goal of determining whether comments are related to return probabilities. Using two complementary text analysis techniques—text regression (CCS) and Linguistic Inquiry and Word Count (LIWC)—we demonstrate that information from the first comment a new user receives improves predictions of new user return. Our work serves as an example of useful predictive features being extracted from very short text comments and illustrates the importance of social feedback on the experiences of new users.

Keywords: new user attrition, churn, Reddit, text analysis, sentiment analysis, online comments, social media, social feedback

---

The first encounter of a new user with an online community can make or break their chances of returning. In a study of large online communities across three languages, 30-70% of users left after their first post and never posted again (Yang et al., 2010). If Reddit is similar, then the average Reddit user is most undecided about their continued use of the website when they make their very first post. It is a make-or-break moment, and many users leave after only one interaction. This drop-off makes sense: new users are, in a sense, testing the waters, seeing what other users are like, and deciding whether Reddit is something into which they want to invest their time.

Historically, churn prediction (predicting which users are most likely to leave, or "churn") has been performed mostly in commercial contexts where money changes hands through purchases or subscription, contexts that usually lack social components. Most prior work, then, is based on factors intrinsic to the individual, like their age, gender,

income, device type, location, or interests. Predicting user return based on extrinsic factors and social feedback is not well-explored. A handful of studies in online contexts have shown that social feedback received in the early stages in a user's lifetime can predict whether they return (Sarkar, 2013; Wang et al., 2013), but studies like these usually use simple measures like views, votes, or scores, while richer social responses like text comments are understudied.

If an enormous proportion of users (estimated at 30-70%, Yang et al., 2010) leave or give up after their first post, then their experiences merit further investigation. Is the social feedback they receive related to their decision to leave? Do "upvotes" tell the whole story, or do comments matter also? Reddit is well-known for the comments and discussions that take place on posts. Can we use modern text analysis techniques to extract predictive features from these comments that can improve our predictions of new user return?

We explore the following research questions in this article:

1. Does the comment feedback of other users predict whether someone will return to Reddit after their first post?
2. If comments are predictive of new user return, which linguistic features of those comments are most powerful as predictors?

To address these research questions, we set for ourselves the cleanest (and most difficult) version of this prediction task, by exploring whether we can predict the return of new users whose post gets exactly one comment. To determine whether the social feedback in the comment itself adds to our predictive ability, we compare the performance of two models: a text-informed model vs. a baseline model that uses other user and post features. Throughout, we illustrate the use of our chosen text analysis methods and explain how they complement each other in extracting text features from our text corpus (Reddit comments) which contain site-specific slang and terms that are difficult to model with an off-the-shelf method alone. Ultimately, we find that comment text features do contain information predictive of new user return, and we discuss our findings' applicability.

WHY REDDIT?

Reddit serves as an interesting case study of the moment people decide whether to join a new community for several reasons. First, its size. Reddit is an enormously popular

website. Use has been on the rise in recent years, nearly doubling in just two years from 11% of US adults in 2019 to 18% of US adults in 2021 (Auxier & Anderson, 2021), with the website boasting 50 million daily unique users across 100,000 active subreddit communities as of January 2021 (*Reddit - Press*, 2021). Reddit's size and popularity alone make it a worthwhile domain of academic study.

Second and relatedly, the volume of posts and comments on Reddit allows for analyses at a vast scale. In this article, we analyze 700,000 pairs of posts and comments using computational means. Such scale means that we have enough data to analyze even relatively "rare" phenomena with large sample size, which is important for quantitative text analysis methods.

Third, the public nature of Reddit, as well as the fact that most users use pseudonyms when they engage with the site and interact with users who they do not know in real life, means that there are fewer outside factors for researchers to control for. On Facebook, for example, online interactions frequently take place in a social context where friends or family members know each other in real life. On Reddit, however, there is very rarely any additional social context unknowable to researchers, besides what is captured in the online interaction itself. Interactions on Reddit, therefore, are more easily conducive to research, because we know that there are unlikely to be outside social factors that we cannot control for.

Fourth and finally, Reddit, its users, and its many subcultures also produce emergent behavior that spills out beyond the website to impact broader society, making it an important topic for academic research. One example is that Reddit was among several social media sites targeted by Russian government-affiliated troll accounts aiming to influence the 2016 US elections (Aleem, 2018). More recently, in 2020 and 2021, users of the subreddit "r/Wallstreetbets" coordinated stock market behavior to bring about the GameStop short squeeze of 2021 (Anand & Pathak, 2022; Mancini et al., 2022). Across time, researchers have identified some Reddit communities as being hotbeds of radicalization (Grover & Mark, 2019; Gaudette et al., 2021), indeed, Reddit was among the four tech companies subpoenaed in 2022 by the "United States House Select Committee to Investigate the January 6th Attack on the United States Capitol" during its investigations

to determine "what steps—if any—social media companies took to prevent their platforms from being breeding grounds for radicalizing people to violence" (Press Release, 2022).

Further, Reddit's impacts on society can be both positive and negative. As one example, in the wake of the COVID-19 pandemic, Reddit was home to much discussion of anti-science and vaccine-hesitant views (Kumar et al., 2022), and Reddit users on political subreddits were found to spread toxic, "rude, disrespectful" pandemic-related news and content more widely than neutral and scientific content (Chipidza, 2021). On the positive side, however, Reddit was a useful forum for patients with long-COVID symptoms to share information and support (Thompson et al., 2022), which in turn provided a useful source of information for medical researchers to explore long-COVID (Sarker & Ge, 2021).

Previous Work on Churn Prediction

Churn prediction, or user/customer return prediction, has a long history among telecom companies (Hung et al., 2006), banks (He et al., 2014), paid television providers (Jamal & Bucklin, 2006), and other subscription services (Coussement & den Poel, 2008). More recently, online gambling companies have engaged in user churn prediction (Coussement & Bock, 2013). Rather little work has been done, though, on social settings, which makes Dror et al.'s (2012) work on new users of Yahoo Answers particularly instructive for this investigation. The researchers use a range of classification techniques from machine learning to make predictions based on only the first week of user activity, (including such features as number and length of questions and answers, and numbers of "thumbs up" vs. "thumbs down" votes on answers) and show that statistically significant signal exists, even in these early users. They found that random forest and logistic regression models performed well in this context, and they report that there are two main sources of signal that are particularly predictive: measures of the user's own activity (in particular, quantity and frequency of contributions) and, of special note for our own work, the amount and type of recognition they received from other users. In the context of Yahoo Answers, recognition takes the form of "thumbs up" and "thumbs down" votes (very similar to Reddit's upvote/downvote system), and the bestowing by the question writer of "best answer" status.

The scope of the work of Dror et al. (2012)—looking at the first week of a user's lifespan and predicting a binary stay/leave outcome—is the closest to ours. However, Yahoo Answers operates as a transactional ask and answer site, which means that users' motivations are different for Reddit. We also use text analysis methods in our work, which offers us a much richer set of predictors and lets us understand more of the nuances of the relationship between comments and new user churn.

METHODS

Regression Methodology

We choose multiple logistic regression as the methodology for both our text-informed model and our baseline model. This is a good fit for our research goals both because our outcome "does this user return?" can be conceptualized as a binary result, and because our main research goals are less interested in perfect prediction than in being able to disentangle the contributions of each individual predictor, which is not possible with most other classification methods. To fight the dual problems of over-fitting to training data and feature selection when faced with many predictors, we also employ L1 regularization (LASSO). Using regularization (with tuning parameters set by cross-validation) and a held-out test set gives us confidence that any success our model has on our test set is not due to overfitting but due to genuine predictive ability.

Data

On Reddit, users can leave comments on posts, and they can also comment on other comments (nested comments), meaning that comment threads can become complicated. The mechanism by which we expect comments to influence post authors (to stay or leave) is if the comments address the post or the author, rather than engaging in discussions with other commenters. If we are to study the relationship between comments and whether a post's author returns to Reddit, we will be dealing with a lot of "noise" if we hope to incorporate long comment threads. In light of this, we consider only those users who received exactly one comment on their very first post. We found that the number of comments received has a strong and positive correlation with the likelihood that a user returns, so this subsetting method controls for that, while also side-stepping the problems of dealing with extremely variable amounts of comment data for each post.

We acknowledge that our choice to study first posts that receive only one comment limits our predictive ability; the number of comments received explains some of the variance in whether or not a first-time author returns. Comparing like-with-like by considering only posts with one comment allows us to study more closely the quality of the comments, which is more interesting to us than their quantity. Looking only at posts with single comments also relieves us of dealing with the different targets and referents and interactions that come with complicated comment structures, where some commenters may be responding to each other rather than to the original post. Looking at posts with only one comment, then, isolates the relationship we're interested in between the text of the comment and the new user's decision about whether to return to Reddit.

We further limit ourselves to look only at text-only posts, rather than images, links, advertisements, or other media content. In this way, we incorporate information about the text of the post as well as of the comment, and can assume that the comment discusses content (text) that the post author wrote, therefore having more of a relationship with their probability of return than third-party content.

The task for this article, then, is to predict whether someone will return to post after their first post, given that their first post was a text post of at least 150 characters that attracted exactly one comment. The dataset we use is the subset of all Reddit posts (up to September 2017) that are the first post by their author and that received exactly one comment. We choose to focus on text posts over 150 characters in length so as not to over-analyze the text of very short posts. Each row in the dataset then, contains a post, its sole comment, and whether or not the post's author would go on to write other posts on the site (coded as a binary variable, with 1 for users who never posted again, and 0 otherwise). There are roughly 700,000 post/comment pairs that occur before the end of September 2017 (the end of our data) that fit these requirements. Our analyses are conducted on random subsets of this dataset. From our 700,000-row dataset, we hold out a testing set of 20,000 post/comment pairs that we do not look at or touch for model training, only using these to compare the performance of our text-informed model and our baseline model.

It should be noted that the data we have does not contain information on when a user account is created, therefore, we use the event of a user's first ever post as a proxy for

their being a "new user", even though it is, of course, possible that much time and activity has elapsed between the creation of the user account and their first post.

Predictors used in Both Models

As was mentioned at the beginning of this article, the framework we used to think about the predictors is that of factors "endogenous" to the post itself (decisions made by the user) and factors "exogenous" to the post (decisions made by others). Endogenous factors include, for example, the subreddit in which the post was made, the word count of the post, the date of the post, and the substance of the post itself (characteristics of its language). Exogenous factors are those that came chronologically after the post was made and originate "outside of" the original author. These include the score of the post (upvotes minus downvotes), the word length of the comment, the length of time between the post and the first comment, and of course, the content of the comment itself.

We wish for our baseline model to reflect, generously, the previous "state of the art" in that it includes every predictive feature we could think of, with the sole exception of our text features. In keeping with this, for the set of predictors to be used in both models, we combined all subreddits below the 100th most popular subreddit into a single "other" category, allowed the year of the post to vary non-linearly with the outcome, and generated some simple text features including word count of posts and comments, ratio of uppercase to lowercase letters in posts and comments, proportion of non-alphabet characters used in posts and comments. The full list of predictive variables used in our baseline model is shown in Table 1.

Predictors for Text-Informed Models

Our next key task is to select text analysis methods to quantify the sentiment and content of the comments on Reddit. Ruling out manual coding from the outset as much too expensive and subjective, we consider two broad categories of modelling approaches: using off-the-shelf dictionary-based methods, and training bespoke models on some labelling of the comments themselves.

When using pre-existing, dictionary-based methods, there are limitations: Reddit-specific words, internet slang, and emojis are unlikely to be well-covered, although they are meaningful and prevalent in the Reddit context. Such dictionaries are limited to unigrams (single words, or 1-grams), which discards the ordering of words, and cannot

capture phrases. This limits our ability to understand linguistic context: dictionary-based methods cannot distinguish between a mean comment like "you suck" and a commiserating comment like "that sucks". They certainly cannot distinguish between a

Table 1
Predictive features generated from raw Reddit data

	Variable Name	**Description**
Endogenous	posts_word_count	Word count of post
	posts_year	Year of post
	posts_weekday	Day of week of post
	posts_hour	Hour of post (from midnight)
	posts_upper_to_lower	Proportion of capital letters used in post
	posts_non_alpha	Proportion of non-alphabet characters used in post
	posts_subreddit	Subreddit the post was made in
Exogenous	posts_score	Score of the post
	comments_score	Score of first comment
	comments_word_count	Word count of first comment
	comments_delay	Time in minutes between post and first comment
	comments_deleted	One if the comment was deleted, zero otherwise.
	comments_upper_to_lower	Proportion of capital letters used in comment
	comments_non_alpha	Proportion of non-alphabet characters used in comment

word's context-dependent meanings, for example, the word "fire" takes on different meanings in each of "she had to fire him", "the house caught on fire", and "that album is fire".

The alternative is to train some sort of model on a labelling of the text. However, such trained models are not able to use information learned from other text corpora or domain-knowledge coded into dictionary methods by linguists and language experts, and would not be able to make predictions for any n-grams that they do not see in their training data.

In this article, we choose to combine the strengths of each type of text method by using one method of each type to generate predictors for our text-informed model. The two methods we employ are Linguistic Inquiry and Word Count (LIWC) and Jia et al.'s text regression (also known as Concise Comparative Summarization, or CCS, Jia et al. 2014).

LIWC Predictors for Text-Informed Model. First created by psychologists Francis and Pennebaker in the 1990s, Linguistic Inquiry and Word Count (LIWC) was built for the task of computerized text analysis (Tausczik & Pennebaker, 2010), and has since been applied to a range of tasks including the analysis of online discourse. LIWC is a dictionary-based method, meaning that words in an input document are compared to dictionaries: collections of words that have been judged to belong to a particular category. The 2015 edition of the software package has around 90 such categories, ranging from simple positive/negative sentiment to emotional and relational words, to grammatical features like usage of pronouns and tense, to words about topics like money or religion, or words that are associated with certain types of thinking, like certainty or causality.

In the vocabulary of Natural Language Processing, LIWC is a bag-of-words model that completely ignores the order of words in the input document. At first blush, this would seem to lose most of the ability to analyze the meaning of the text, but countless researchers and studies have shown that bag-of-words representations can still yield useful insights. LIWC first stems the words as part of its pre-processing, trimming different conjugations like "tempting", "temptress", "tempted", and "temptation" down to "tempt-". Categories are then defined by the membership of a few dozen or a few hundred words considered to represent that construct. LIWC is a frequency-based method; essentially, it counts all the words belonging to a given category, divides that by the number of words in the document, and returns a series of percentages. For example, it might conclude that a given document was made up of 7.5% pronouns and 4.2% positive emotion words, and so on. In total, there are over 6,000 words and word stems in LIWC's dictionaries. Each of the dimensions is reported as a number between 0 and 100, representing the percentage of words in the document that were flagged as belonging to that category. Because our comments vary wildly in length, we expect large variances here.

Because some LIWC variables are hierarchically defined—words associated with sadness are categorized as negative emotion words, which all are also counted as overall affect words—multicollinearity in our predictors became a concern. We therefore proceed with a dataset in which 13 "parent" LIWC categories have been dropped for both posts and comments, greatly reducing the amount of multi-collinearity in the dataset (as measured by variance inflation factors) and leaving us with 170 remaining LIWC predictors.

Text Regression Predictors for Text-Informed Model. One thing that's especially exciting about text regression is that it can find phrases and patterns for which we don't explicitly look. We don't know in advance all the words and phrases that might be associated with comments that encourage Redditors to stay or leave.

Text regression, also called Concise Comparative Summarization (CCS) is a method designed for the analysis of labelled text corpora (Jia et al., 2014; Miratrix & Ackerman, 2016). One can think of it as a means of regressing a binary classification on text data, hence "text regression". The name "Concise Comparative Summarization" sheds even more light. The technique uncovers the words and phrases that distinguish between two sets of documents, usually a baseline and a set of interest, which is the comparison element.

One strength of text regression is its ability to include arbitrarily long phrases. This allows us to retain some of the ordering information lost when simple bag-of-words methods like LIWC are used. We used the R implementation provided by the *textreg* package (Miratrix, 2017) in our work.

We trained our text regression model directly on the training set of users who did and did not return to Reddit after their first post, and then used the list of n-grams produced by text regression as predictors in our text-informed model. Text regression was essentially used as part of our feature engineering process, as the *textreg* package did not allow for the use of other covariates. We chose to run text regression only on the text of the comments (not the posts), allowing us to directly address the role of the qualitative feedback users receive on their very first posts. When training the text regression model, we used a random training set of 100,000 post/comment pairs (larger than the random training set of 20,000 we eventually used for the logistic regression). The phrases we are

looking for are extremely rare, as text data is sparse and each comment is short, so we wanted to cast a wide net.

There are several hyper-parameters associated with text regression. We capped the length of the phrases allowed to ten words, specified that we didn't want any phrases to be returned that didn't exist at least five times in the training data, didn't allow for wild-card gaps in the phrases (primarily to limit computational costs), told the model that we were interested in *both* "positive" (predictive of users leaving) and "negative" (predictive of users returning) phrases, chose to look only at the presence or absence of phrases rather than their counts, and set the token type to be words rather than characters.

The two tuning parameters that are most important are implemented in the *textreg* package as C and Lq. C is the main tuning parameter that regularizes the regression. Larger Cs led to smaller phrase lists, because they make it more difficult for a phrase to be selected. We used a C of 8.3, chosen via cross-validation on unseen data.

Lq represents the q value for the L^q rescaling of terms. Rescaling the words by their frequency of appearance is part of why text regression doesn't require an explicit stop-word list, which is traditionally one of the larger challenges of natural language processing. Bigger values mean more common phrases can appear, while smaller values of Lq favor rarer phrases that more powerfully distinguish between the two groups. We chose an Lq value of 1.75 by hand, as it seemed to produce phrases that were interpretable and meaningful without being too rare in the data. We preferred for our list to be perhaps slightly too long than too short—feature selection through regularization should weed out phrases that become unimportant in the presence of other predictors.

Outcome Variable & Defining User "Death"

Whether or not someone has left Reddit—which we refer to as user "death"—is not available to us in the dataset. Indeed, it is not available to Reddit itself unless the user explicitly requests to delete their account. What to do then? If someone hasn't posted in, for example, a year, we can fairly safely infer that they have left the site, even though a user returning after more than a year of inactivity is not impossible. However, saying that someone who hasn't posted in a month has left Reddit is probably too short a time window. Something in between these two lengths is probably reasonable.

Other researchers have used different cut-offs for user "death" in online contexts, and there is no agreed-upon method here. Perhaps the most helpful approach is documented by Yang et al. (2010), who declare an account to have died after a period of inactivity exceeding 100 days on the three Question-and-Answer websites they study. The researchers attempt to justify this cut-off by stating that, in their data, fewer than 30% of users had more than 100 days between posts. This seems like an awfully large number of living users to declare dead. Crucially, though, they performed a sensitivity analysis using cut-offs of 50 days and 150 days to see if their cut-off decision had an effect on their model results, and they found that changing the definition of death didn't change their conclusions. This is reassuring.

In our own research, we set about analyzing the appropriateness of different cut-off rules by looking at the data first. We took a sample of 20,000 Reddit accounts from the full Reddit dataset and retrieved the timestamps of every comment associated with these users (over 17 million such comments). Grouping comments by user and taking the average gap between comments for each user, we found that the 99th percentile of user-level averages was 121 days of delay. This is encouraging, revealing that past researchers have been in the right "ballpark". While users make posts and comments at different rates, we feel that this is a good rule of thumb to use for declaring an account to be officially "dead".

Because our potential dataset encompassed the whole of Reddit, with posts dating back well over ten years, all this discussion of cut-offs only really affects our analysis of newcomers to the site towards the end of our dataset (we analyzed data up to the end of September 2017). We decided to consider our new users as returning if they made another post before the end of the time window, and "dead" only if they never reappear in our data. For example, if a brand-new user made their first post in 2009 and hasn't posted since, does it matter at which exact moment we can declare them to be "dead"? But if someone made their very first post within the final month, we cannot really be sure whether they might still return or instead have already decided to quit. Ultimately, in the dataset used in this article, we chose to drop from our analysis any users whose first post was made after the beginning of June 2017, to allow for around 121 days of observation to see if they returned. In our test set of 20,000, 53.46% of post authors left while the rest returned.

This is the accuracy rate that would be achieved by a completely stupid classifier that guessed that every single new user left, and will be the benchmark against which our baseline and text-enhanced prediction models will be compared.

Model Evaluation

We set out in advance the metrics by which we would compare the models constructed in this article. We use classification rates, confusion matrices, and AUCs (Areas Under the Curves, with DeLong's significance test at the 0.05 level). DeLong's test is a non-parametric test to compare the area under two or more ROC curves (DeLong et al., 1988). We use DeLong's test as implemented in the roc.test function in the R package pROC (Robin et al., 2011). Our RQ1 is "Does the comment feedback of other users predict whether someone will return to Reddit after their first post?" These model evaluation methods will allow us to address whether the predictive ability of the models with text features generated from the user comments will significantly outperform the baseline model that does not use these comment features.

RESULTS

Baseline Model

To set our baseline, we predict whether someone will leave after their first post on Reddit, *without* incorporating our text analytic predictors that we hope will shed light on the role of comments and feedback. The value of lambda chosen by cross-validation that yielded the minimum mean cross-validated error for these predictors was 0.0003, which is very light regularization. We truncate the list here to remove the 10 different subreddit terms and omit year variables for length. A dot represents a coefficient that was shrunk to zero by regularization. Positive coefficients indicate an association with leaving Reddit, negative coefficients with returning.

Table 2
Estimated regression coefficients for the baseline model (3 s.f.)

Intercept	.
posts_score	-0.0249
comments_score	-0.0151
comments_deletedTRUE	0.0445
comments_word_count	0.000674
comments_upper_to_lower	0.0788
comments_non_alpha	0.0991
posts_word_count	0.000166
posts_upper_to_lower	0.329
posts_non_alpha	0.399
comments_delay	0.00000418
posts_weekdayMonday	-0.0308
posts_weekdayTuesday	-0.0343
posts_weekdayWednesday	-0.0445
posts_weekdayThursday	-0.0185
posts_weekdaySaturday	.
posts_weekdaySunday	.
posts_hour	-0.00146

Higher post scores are associated with a decreased chance of leaving. Users who stayed often received comments with higher scores. If the comment on a post was deleted (either the commenter or a moderator could have deleted it) that decreases the predicted chances of staying (deletion is often a sign that the comment was unsavory or spam). Posts on weekdays further away from the comparison day, Friday, are associated with a higher chance of returning. Interpreting this through the lens of odds ratios gives us a way to describe the magnitude of this effect on our predictions: This model predicts that the odds of a user leaving whose first post had a deleted comment to be $e^{0.0445} = 1.046$ times the odds of someone who had a comment that was not deleted. A roughly 5% increase in the predicted odds of leaving seems reasonable. The *comments_delay* variable is coded in minutes, over which the change in predicted odds is very small, but if your post only gets its first comment after a full week, your predicted odds of leaving also increase by around 4%: $e^{60\times24\times7\times0.00000418} = 1.043$ times the estimated odds with an immediate comment. Compared to these coefficients, the scores of the post and comment seem especially important. An author whose post received ten upvotes has predicted odds of leaving $e^{10\times-0.0249} = 0.78$ times their odds if they had received no upvotes, and a comment with ten upvotes $e^{10\times-0.015} = 0.86$ times that of a comment with a score of zero.

Curiously, posts and comments with longer word counts are associated with an increased chance of leaving, which did not align with our expectations. We remain unsure why this is the case. The classification rate for the predictions on the test set made by this model was 0.5874, with a 95% confidence interval of (0.5805, 0.5942) (here, the confidence interval represents how much our classification rate would move if we resampled 20,000 training users and re-ran repeatedly). Looking only at a first post and a single comment, we are able to predict better than by 50/50 chance, and better also than predicting every user to leave (53.46% of test set users leave) but this limited information isn't enough for our model to do an amazing job of predicting whether new user churn. The fact that we can predict at all is somewhat remarkable given the limited information we allowed ourselves about each person and their first interaction on Reddit. The confusion matrix for these predictions is reproduced in Table 3, expressed as percentages:

Table 3
Confusion matrix for baseline model

		Truth	
Prediction		Returned	Left
	Returned	22.335%	17.000%
	Left	24.260%	36.405%
	(Total)	(46.595%)	(53.405%)

To see whether non-linearity in any of our predictors affected our model results, we also fit a Generalized Additive Model (GAM) with basis splines on all of the numeric predictors (but without any regularization). A GAM performed similarly in predictions on this test set, with a classification rate of 0.58555, very similar to, and within the confidence interval of, our previous classification rate. This indicates that the regularized linear model is working well, and that the non-linearity of the predictors doesn't seem to be too important for prediction purposes. It also suggests that the regularization may not be needed. However, because our text features introduce several hundred new predictors to the model and we wish to both use the same approach in both models and err on the side of caution, we continue using regularization.

Because we are especially interested in the relationship between comments received and user return, we now ask whether the addition of text predictors will improve our predictive accuracy.

Model with Text Predictors

In this section, we add the text predictors to the baseline model to assess how well they improve our predictive ability. Does the content of the comments themselves give us more information about whether a new user might stay or leave?

The regularization penalty λ that resulted in the smallest mean squared error here was 0.002, very similar to the baseline model. Because this model now contains hundreds of predictors, only a truncated list is presented here, and the baseline predictors all look relatively similar in this model.

Table 4
Selected estimated LIWC coefficients for comment features in the text-informed model (3 s.f.)

2nd Person Pronouns	0.00593
Negations	0.0002.89
Interrogatives	0.00273
Question Marks	0.000780
Negative Emotion: Anger	-0.00939
Social Processes: Friends	-0.0232
Swear Words	0.00744
Assent	-0.00344
Exclamation Marks	-0.00267

Looking down the list of LIWC predictors (selected LIWC coefficients in Table 4), there are certainly some relationships that make sense. Increased use of swear words in comments is associated with an increased chance that the recipient leaves Reddit, as are words associated with negating. Comments containing words about friendship and assent are associated with staying. Curiously, higher rates of second person pronouns—probably directly addressing the poster—are associated with an increased probability of leaving.

Our findings from the LIWC predictors are limited in what they tell us about the nature of the feedback received. What does a higher concentration of exclamation marks in the comments on users who stayed, for instance, really *mean*? And given that angry sentiment words are more common among comments on those who stayed, is that anger towards the post author or shared anger towards some third party or object? Why are question marks and interrogatory words predictive of leaving? These are the types of questions LIWC cannot answer. While these substantive questions may generate fruitful

lines of research suited for qualitative methodology, they are beyond the scope of the research questions in this paper.

A few themes emerge in the text regression n-grams (selected text regression coefficients in Table 5). The first is that there are many phrases that are possibly to do with spam, such as “car insurance”, “click here”, “compare quotes”, “credit”, “insurance”, “lawyer”, “online free”, and “torrent”. This is fascinating to discover; we could not have specified such things in advance, but the text regression method was able to extract them. The fact that these variables are associated with an increased likelihood that the recipient leaves the website makes sense: it’s annoying to receive spam comments. They might signal to new users that the site (or subreddit) is poorly managed and moderated. Additionally, in the context of posts with single comments, spam comments also mean the lack of an interaction with an actual human which might in itself be discouraging.

Table 5
Selected estimated text regression coefficients for the text-informed model (3 s.f.)

“car insurance”	1.04
“click here”	0.289
“compare quotes”	2.22
“credit”	0.528
“insurance”	1.20
“lawyer”	0.327
“online free”	1.03
“torrent”	1.53
“please edit this”	0.433
“please message the mods”	-0.324
“require that you”	-0.822
“added you”	-0.0325
“request sent”	-1.42
“another job”	0.189
“depressed”	0.238
“depression”	0.134
“doctor”	0.0308
“sorry you”	0.119
“therapy”	0.596

Second, a group of phrases that stands out are formal phrases that appear to be moderator rebukes, including phrases like “please edit this”, “require that you”, and “please message the mods”. These have differing relationships with user return, some with

positive coefficients and some with negative coefficients. One possible explanation for this is that some subreddits use automatic moderating bots for certain rule infractions, so some of these comments might have come from auto-mods. This also means that there hasn't been a real human interaction. Receiving feedback from an actual moderator might be discouraging for some, but on the other hand, it might be a sign that the community is active. If an important and well-known user (which moderators tend to be) interacts with a newcomer, this might be exciting. It's difficult to say much more based solely on the regression coefficients, but this would be a fruitful avenue for future research: How does interaction with a moderator affect survival probabilities?

Third, two social phrases—"added you" and "request sent"—indicative of an online relationship forming between the commenter and the poster, are associated with a higher probability of the poster returning.

Fourth, we see some n-grams that we interpret to be responses to difficult situations or problems shared by the poster. Words and phrases such as "another job", "depressed", "depression", "doctor", "sorry you", and "therapy" are *all* associated with an increased chance of leaving. There are two likely mechanisms here. One is that this first-time poster was troubled and unlikely to continue on Reddit, regardless of the comment they received. This is plausible. The other possibility, perhaps more compelling, is that this person created a "throwaway" account specifically for the purposes of sharing this personal problem or personal story. This is a well-established behavior on Reddit, and it could explain this trend (Ammari et al., 2019; De Choudhury & De, 2014).

This text-informed model achieved a classification accuracy of 0.6023, with a 95% confidence interval of (0.5955, 0.6091). This is a small but significant (and statistically significant) improvement over the previous model, which had an accuracy of 0.5874. The lower end of our confidence interval doesn't include the baseline accuracy, which is evidence that these text predictors together do indeed improve our predictions and add to our understanding of user retention.

In this confusion matrix (Table 6), we see an interesting behavior. The number of false positives (users predicted to leave who actually stayed) dropped by 739 of our test set of 20,000. At the same time, though, the model now misclassified an additional 441 users as returning when they really did not, an increase in the number of false negatives. These

additional text predictors do improve predictions overall, but at the cost of reduced performance in some areas.

Table 6
Confusion matrix for the text-informed model, with differences from baseline model in parentheses

		Truth	
Prediction		Returned	Left
	Returned	26.030% (+3.695%)	19.205% (+2.205%)
	Left	20.565% (-3.695%)	34.200% (-2.205%)
	(Total)	(46.595%)	(53.405%)

The ROC curves and AUC (Area Under the Curve) metrics for both this final model and the baseline model are shown in Figure 1. We see that the new model doesn't appear to completely dominate the baseline model, as there are some points towards the extremes where the baseline model appears to perform better. The overall AUC has increased slightly from the baseline to the text-informed model with the addition of the text features, moving from 0.635 to 0.645.

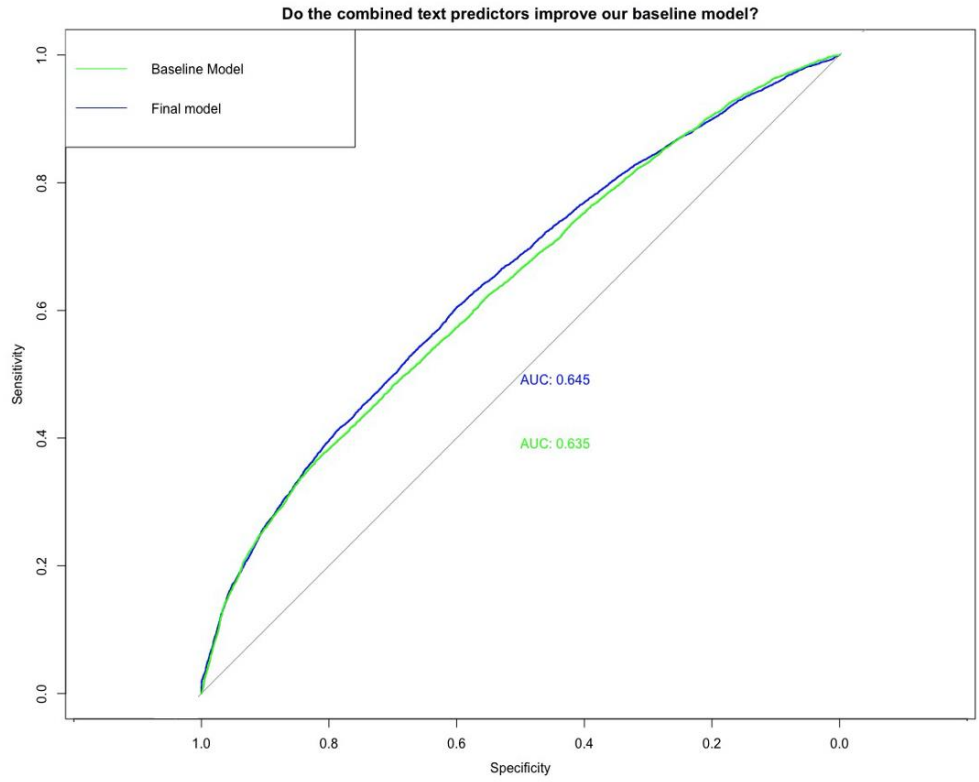

Figure 1. The Receiver Operating Characteristic (ROC) curves for the two models, mapping true positive rate against false positive rate

In a test for the significance of this difference in the AUCs, DeLong's test gives a p-value of 0.0009084, rather strong evidence in favor of a statistically significant difference, even if it is not a large one.

Comparing Text Methods

Future researchers may be interested in using a combination of text analysis methods as we have done, but is there compelling evidence to do so? Does using the two different text analysis methods *together* add information that is more than either could provide by themselves or are they too similar?

Our evidence is not clear on this front. We re-fit the model using the baseline predictors plus either the LIWC or the text regression predictors, and then we compared these two models against our baseline and our combined, text-informed model. In terms of classification rates, a model without text regression predictors achieved an accuracy of 0.5991, higher than the baseline 0.5874, and a model without LIWC predictors achieved an accuracy of 0.5950, also higher than the baseline. The lower end of the confidence interval for the final model's accuracy excludes all three of these, which suggests that the use of *both* types of text predictors is statistically significantly more accurate than just one. On the other hand, when using DeLong's test to compare the final model to each of the component models, we find a p-value of 0.1102 for the final model over the model with LIWC predictors (weak evidence), and a p-value of 0.8255 for the final model over the model with just text regression predictors (no evidence at all).

In our very particular context and application, we can't say for sure that the use of both types of text predictors was significantly better than just one, though some evidence points in that direction. We think that future researchers would be wise to incorporate more than one method of text analysis in similar work wherever possible. The LIWC method has the strength of not requiring any sort of labels as it is an unsupervised method, but if labelling is possible, text regression provides far more interpretable predictors. Broadly speaking, we can consider the LIWC predictors to be getting at *how* the comments are phrased (with a range of measures for grammar, punctuation, and tone) whereas the text regression predictors do a better job of describing *what* was said. Both seem useful in predicting user churn, and in other prediction contexts.

DISCUSSION

In this article, we explored features of Reddit comments that are helpful in predicting whether a first-time poster on Reddit will return to post again. Beginning with simple predictors drawn from the Reddit API, we then looked at the additional benefit of adding text analytic features. We found that both predictors drawn from LIWC dimensions and predictive phrases found through text regression were statistically significant in adding to our predictive ability and that our final model with all predictors performed best. Gains in predictive accuracy were modest, but this is all we should have hoped for, given the huge array of factors that influence people's decisions as to whether to post again that *aren't* tied to the feedback they received on their first post, and given the extremely narrow slice of data we allowed ourselves (a user's first post, so long as they received only one comment).

In addressing the research questions we posed at the beginning, we can now say that yes, the feedback of other users predicts whether someone will return to Reddit after their first post. This feedback comes in many forms: the presence of a comment and the delay before that comment was posted, the score the post received, and the content and style of the comment text itself. We used two methods to extract textual features of comments that are predictive of user churn, and we found that many of them were significant, though not all were interpretable. As for the linguistic features of the comments, we found through text regression many n-grams that were predictive of staying or leaving, and found through LIWC many stylistic and grammatical features that differed on both the posts and comments between users who stayed or left.

Methodological Contributions

The main methodological contributions of this article are two-fold. Firstly, the application of these text analysis techniques to a dataset of this size and type is still an emerging art form, and it brings with it many challenges, such as documents of uneven lengths, domain-specific vocabulary, and the difficulty of selecting an appropriate number of parameters in a principled way. We hope that this article is encouraging as a demonstration and proof-of-concept to future researchers of the potential for rich text predictors to be incorporated in their work, even in traditional regression set-ups, and that these text predictors can add to both the accuracy of our models and our qualitative

understanding of the context. We also hope that our choices in implementation can serve as an interesting example for future research. Previous research using LIWC in online contexts is often limited to looking only at the positivity and negativity of sentiment, just two predictors of the dozens that are possible. For some contexts, using an even richer selection of text predictors might offer more insight.

This work also represents a somewhat novel use of text regression, in that we have demonstrated that it can be used as part of a feature engineering process before being combined into a larger prediction model. The added predictive ability and insights gleaned from the words and phrases produced by text regression proved valuable, and they were in many ways more interpretable and shed more light on the domain than the LIWC predictors.

Reddit-specific Contributions

There are also some Reddit domain-specific contributions in this article. Showing that the feedback received in comments is predictive of users staying or leaving Reddit after their first post demonstrates that some portion of the huge amount of user drop-off after first posts can be explained by the feedback they received and not just by factors intrinsic to the user. While this is an intuitive finding, it is exciting to be able to demonstrate statistically. The fact that the content of the comments, not just the post's score, influences users doesn't go without saying.

Limitations of this Work

The conclusions drawn in this article cannot be extended too widely, given that we chose to take a very specific subset of the data with which to work. Though similar trends likely persist in users whose first post received more than one comment—and not just among people's first posts but also later posts—these effects might be smaller or otherwise different. We also chose to examine posts as the main unit of analysis rather than comments. But comments are far more numerous on Reddit than posts and our findings do not necessarily transfer to them.

We also saw evidence suggesting spam and automatically generated comments in our data. We don't know how much these influenced our results, though it would be difficult to remove these from our data set with any precision.

We should acknowledge that this way of framing the problem (just looking at new user *posting* behavior) ignores new user commenting behavior. It's possible that some of our "new" users had already made comments before their "first post" on the site, and it's possible also that users who never make another post remain active as commenters on the site. However, any conclusions we can draw, though they may in a few cases not truly address whether a brand-new user left the site entirely, do at least address the question of whether someone's first experience making a post is related to their decision to make more posts in the future. We operationalized the idea of "posts in the future" without any restrictions on the nature of the posts. Future posts can be links or text posts of under 150 characters and still allow the user to be considered a returner. Other, more complicated choices, including future commenting behavior may be of interest to others.

Opportunities for Future Research

As text analysis methods continue to develop, there are more opportunities for future researchers to apply them to other large datasets like Reddit.

Others may also want to investigate whether different modelling approaches might be more effective in prediction in this context or might do a better job of incorporating information extracted from the text. We recognize that our decision to use regularized logistic regression was one of myriad classification methods we could have used, and we so cannot forecast whether other techniques might produce similar or better results.

Finally, there might be room to apply some quasi-experimental methods of the sort that are used to draw causal conclusions from observational data such as this. Our own research design cannot demonstrate whether the text of the comment *caused* someone to leave. Future work using either quasi-experimental methods or more isolated laboratory experiments could prove exciting.

References

Aleem, Z. (2018, April 11). Reddit just shut down nearly 1,000 Russian troll accounts. Vox. https://www.vox.com/world/2018/4/11/17224294/reddit-russia-internet-research-agency

Ammari, T., Schoenebeck, S., & 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), 135:1–135:30. https://doi.org/10.1145/3359237

Anand, A., & Pathak, J. (2022). The role of Reddit in the GameStop short squeeze. *Economics Letters*, *211*, 110249. https://doi.org/10.1016/j.econlet.2021.110249

Auxier, B., & Anderson, M. (2021, April 7). Social Media Use in 2021. Pew Research Center. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Chipidza, W. (2021). The effect of toxicity on COVID-19 news network formation in political subcommunities on Reddit: An affiliation network approach. *International Journal of Information Management*, *61*, 102397. https://doi.org/10.1016/j.ijinfomgt.2021.102397

Coussement, K., & Bock, K. W. D. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, 66(9):1629–1636. https://doi.org/10.1016/j.jbusres.2012.12.008

Coussement, K., & den Poel, D. V. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1):313– 327. https://doi.org/10.1016/j.eswa.2006.09.038

De Choudhury, M., & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*. https://www.sushovan.de/research/reddit-icwsm.pdf

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3). https://www.jstor.org/stable/2531595

Dror, G., Pelleg, D., Rokhlenko, O., & Szpektor, I. (2012). Churn prediction in new users of Yahoo! answers. *Proceedings of the 21st International Conference on World Wide Web*, pages 829–834. https://doi.org/10.1145/2187980.2188207

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2021). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society*, *23*(12), 3491–3508. https://doi.org/10.1177/1461444820958123

Grover, T., & Mark, G. (2019). Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 193–204. https://doi.org/10.1609/icwsm.v13i01.3221

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia Computer Science*, 31:423–430. https://doi.org/10.1016/j.procs.2014.05.286

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3), 515– 524. https://doi.org/10.1016/j.eswa.2005.09.080

Jamal, Z., & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. *Journal of Interactive Marketing*, 20(3),16–29. https://doi.org/10.1002/dir.20064

Jia, J., Miratrix, L., Yu, B., Gawalt, B., El Ghaoui, L., Barnesmoore, L., & Clavier, S. (2014). Concise comparative summaries (CCS) of large text corpora with a human experiment. *Annals Of Applied Statistics*, 8(1):499–529. https://projecteuclid.org/euclid.aoas/1396966296

Kumar, N., Corpus, I., Hans, M., Harle, N., Yang, N., McDonald, C., Sakai, S. N., Janmohamed, K., Chen, K., Altice, F. L., Tang, W., Schwartz, J. L., Jones-Jang, S. M., Saha, K., Memon, S. A., Bauch, C. T., Choudhury, M. D., Papakyriakopoulos, O., Tucker, J. D., … Omer, S. (2022). COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: An observational study on reddit. *BMC Public Health*, *22*(1), 446. https://doi.org/10.1186/s12889-022-12824-7

Mancini, A., Desiderio, A., Di Clemente, R., & Cimini, G. (2022). Self-induced consensus of Reddit users to characterise the GameStop short squeeze. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-17925-2

Miratrix, L. (2017). *textreg: n-Gram Text Regression, aka Concise Comparative Summarization*. R package version 0.1.4. https://cran.r-project.org/web/packages/textreg/index.html

Miratrix, L. W., & Ackerman, R. (2016). Conducting sparse feature selection on arbitrarily long phrases in text corpora with a focus on interpretability. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. https://doi.org/10.1002/sam.11323

Press Release. (2022). *Select Committee Subpoenas Social Media Companies for Records Related to January 6th Attack*. https://january6th.house.gov/news/press-releases/select-committee-subpoenas-social-media-companies-records-related-january-6th

Reddit—Press. (2021, January). https://www.redditinc.com/press

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77. https://doi.org/10.1186/1471-2105-12-77

Sarkar, C. (2013). *The effects of participation and feedback received on the length of time members in online communities remain active*. PhD thesis, Michigan State University.

Sarker, A., & Ge, Y. (2021). Mining long-COVID symptoms from Reddit: Characterizing post-COVID syndrome from patient reports. JAMIA Open, 4(3), ooab075. https://doi.org/10.1093/jamiaopen/ooab075

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. https://doi.org/10.1177/0261927X09351676

Thompson, C. M., Rhidenour, K. B., Blackburn, K. G., Barrett, A. K., & Babu, S. (2022). Using crowdsourced medicine to manage uncertainty on Reddit: The case of COVID-19 long-haulers. Patient Education and Counseling, 105(2), 322–330. https://doi.org/10.1016/j.pec.2021.07.011

Wang, T., Wang, K., Erlandsson, F., Wu, S., & Faris, R. (2013). The influence of feedback with different opinions on continued user participation in online newsgroups.

ASONAM '13, pages 388–395. ACM and IEEE. https://doi.org/10.1145/2492517.2492555

Yang, J., Wei, X., Ackerman, M. S., & Adamic, L. A. (2010). Activity lifespan: An analysis of user survival patterns in online. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1466

Funding and Acknowledgements

The author declares no funding sources or conflicts of interest.

Online Connections

Twitter: @Emma_Klugman
Website: emmaklugman.github.io
LinkedIn: @emmaklugman