IS YOUR ADVICE UNETHICAL?

SCOTT ATKINSON

1. Introduction

Ethics is an important facet of civilization, and a community's ability to discern between ethical and unethical behavior is critical for a healthy society. In any business, school, or community, ethical communication and behavior is of high importance. For many situations, it is not difficult to decide if an action is ethical or not, but there are also many scenarios lying in an ethical gray area. The growing scale of text communication makes moderating social media sites, internal business communication platforms, and other online forums impossible to do manually–calling for an automated approach to detect unethical behavior.

This project applies machine learning techniques to try to statistically detect ethical advice and answer the question: **What makes a piece of advice unethical?** We will use data from two subreddits on the Reddit website: LifeProTips and UnethicalLifeProTips. For most of the posts in these subreddits, a typical well-adjusted adult will be able to accurately classify them into their proper subreddits. That is, there are not that many "ethically gray"posts in our dataset. So it is a reasonable expectation that a binary classification model can be trained to flag unethical advice.

Our analysis shows some general trends in the nature of the posts coming from the two subreddits. Advice coming from LifeProTips is generally aimed toward self-improvement. That is, most of the tips involve actions an individual can take to improve the quality of their life. These actions typically do not depend on the cooperation or participation of any other individual. For example:

- "hold both ends of the tube and run it over the edge of the sink to push toothpaste to the top - you'll get almost every last bit with almost no effort"

On the other hand, the pieces of advice coming from UnethicalLifeProTips generally involve some sort of dishonest behavior an individual can adopt with the goal of gaining some sort of advantage (financial, social, emotional, or otherwise) from someone else, often at their (the other's) expense. For example:

- "The best way to hang up on someone is in the middle of your own sentence. That way they never suspect you of hanging up on them."

So the labels for this binary classification problem can be set to "self-improvement"or "dishonesty with the purpose of gaining from others."When one really thinks about it, one could conclude that the purpose of all dishonesty is to gain from someone else.

In §2, we discuss the data sources, data preparation, and data analysis. In particular, we examine some expected word frequencies, compare reading levels across the subreddits, and cluster the posts to find some common topics. We also assess the most predictive words for each label with a multinomial naive Bayes analysis. In §3 we assess and select our classification model and the proper threshold for the intended use of our model. In §4 we draw our conclusions on our findings and discuss potential improvements for the model.

2. Data

2.1. **Data sources.** The data for this project are obtained from two subreddits on the online forum site www.reddit.com: LifeProTips and UnethicalLifeProTips. The LifeProTips subreddit contains user-generated content in the form of advice, hints, and tips for various things. The UnethicalLifeProTips subreddit contains similar content with the difference being that the tips are unethical, or at best, in an ethical gray area. The data was obtained by scraping the most recent 5000 posts from each subreddit in early January 2021 using the `pushshift` API. The posts collected run the spectrum from making you laugh to making you cringe. We will see samples from each subreddit throughout the report.

2.2. **Data preparation and analysis.** The data from each subreddit were assembled into their own respective dataframes with the following attributes: `subreddit, title, id, created_utc, score, num_comments, selftext`. For training purposes, we concatenate the two dataframes into a single dataframe. The `subreddit` column serves as our label/target column. In the `title` and `selftext` columns, we pass all letters to lowercase and remove all punctuation. We also remove all instances of the abbreviations "lpt"and "ulpt"along

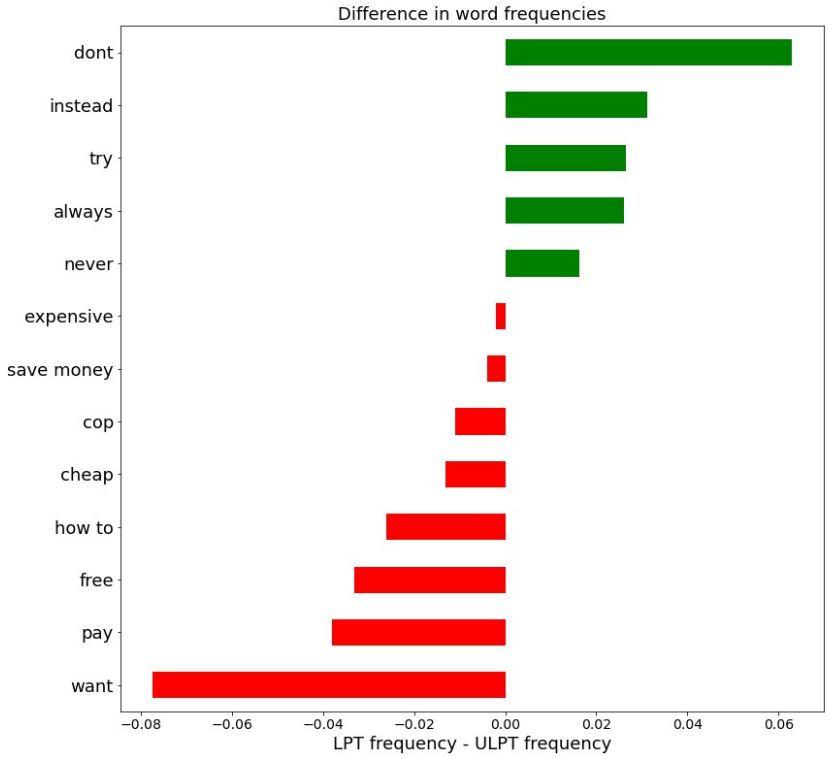

FIGURE 1. Difference in frequencies of words in each subreddit

TABLE 1. Flesch-Kincaid grade levels for `title`

subreddit	mean	std	Q1	median	Q3
LifeProTips	7.0028	3.5783	4.5	6.8	9.1
UnethicalLifeProTips	6.3536	3.8397	3.8	6.0	8.3

with any appearance of "unethical." After cleaning, LifeProTips has 4945 entries, and UnethicalLifeProTips has 4926 entries in the dataset.

After a preliminary look at some entries for each subreddit, we form a list of some words that could be predictive for each and compare their frequencies for each subreddit. Figure 1 provides a visualization for some of the differences between the word frequencies. From a preliminary check of the two subreddits, LifeProTips appears has many pieces of advice telling you not to do one thing and to do another thing instead, and UnethicalLifeProTips evidently has many posts about how to pay less or no money for goods or services.

We next consider the reading levels of the posts from each subreddit. We used the `flesch_kincaid_grade` reading level function from the `textstat` module to evaluate the reading levels of the *unprocessed* title post. The Flesch-Kincaid grade level is given by the following formula.

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

The output returned is meant to roughly align with U.S. grade levels. We retain the punctuation for the reading level computation because sentence count needs to be taken into account. We pass this value to the `title_reading_level` column of the dataframe. The summary data returned is included in Table 1, and Figure 2 displays the box plots for reading level of `title` by subreddit.

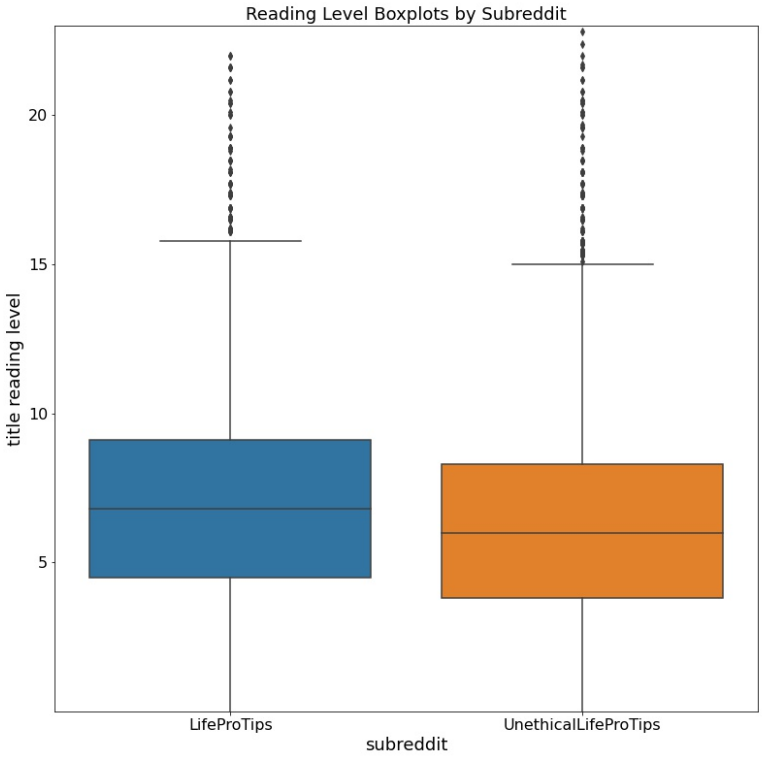

FIGURE 2. Box plots displaying the distribution of reading levels for the titles grouped by subreddit

2.3. **Clustering topics.** We apply the unsupervised learning technique of clustering to the `title` column of each subreddit. We first vectorize `title` and then apply SciKit-Learn's `LatentDirichletAllocation` (LDA) algorithm to the vectorized entries. LDA returns distributions for a prescribed number of topics/clusters, and then each post receives a score indicating how likely it is to belong to each of the topics. Using 7 or 8 clusters makes some topics from each subreddit readily apparent. The topics identified here help give more context to what makes a piece of advice ethical.

2.3.1. *LifeProTips topics.* We first list some of the major topics appearing from applying LDA to the LifeProTips `title` column. Note that these topics and their examples align with the observation that posts from LifeProTips are generally meant for self-improvement with no dependence on the cooperation of others. We list these topics in no particular order.

(1) **Resolutions:** As mentioned above, the data were scraped from Reddit in early January 2021, so there were understandably a large number of recent posts addressing New Year's resolutions. Here are some samples from this topic:
   - "Don't wait for the first day of the new year to start or change something. TODAY is the day to start, no matter where it falls on the calendar!"
   - "When you make New Year's Resolutions, make a plan for when you're tempted and for when you fall short."

(2) **Cooking/kitchen tips:** There was significant representation of tips and advice regarding cooking and preparing food. Some notable samples are:
   - "Don't use the microwave to heat up pasta from the fridge, but use a frying pan and a little bit of water"
   - "If youre a regular iced coffee drinker, freeze an ice tray of milk instead of water."

(3) **Money:** Tips about money management were often clustered together. Samples:
   - "Treat your savings contributions like they're bills you need to pay"

- "Whether you make a lot of money or very little money, you need to have a budget"

(4) **Gift giving:** Again, because the data were scraped near the holidays, there were many posts referring to Christmas gifts. Some samples include:
- "Keep an ongoing list of potential Christmas presents for your loved ones throughout the whole year. Every time they mention something they'd like to have, write it down"
- "When gifting someone a book, always add a note inside. It makes it much more personal and memorable."

(5) **Mental health:** This is not a surprising topic to have received attention in the LPT subreddit. Some examples of posts are the following:
- "Love Your Body"
- "Do not react to anything overwhelmingly the same day it happens. Give yourself a nights sleep and attack it the next day. It chemically allows your brain to process it properly without the flood of emotions and confusion."

(6) **Cleaning:** Many tips surrounded the topic of cleaning. Here are a few examples:
- "When cleaning the house, if you're bringing something from one room to another, bring back something that belongs to the room you were in."
- "Keep a bottle of surface cleaner and a rag in your shower, and clean your shower while you shower."

(7) **Online behavior:** Another common theme is advice regarding online behavior: managing online profiles, accounts, passwords, etc. Samples:
- "have 3 passwords: one for your main mail account, one for websites with your card info, one for the other"
- "If you cannot put a tape against your laptop camera, then try disabling the camera in device controller."

2.3.2. *UnethicalLifeProTips topics.* Note that these topics and examples support the observation that the posts from UnethicalLifeProTips suggest dishonest behavior that takes advantage of others.

(1) **Avoid ads/paywalls:** This topic includes tips on how to avoid ads or paywalls on various websites. This topic is a more passive form of dishonesty. Samples:
- "If you get stuck behind a paywall for a news article the pay wall can usually be removed by using control + shift + I and deleting the pay wall website element."
- "Want to use Youtube to listen to music, but don't want to listen to ads? Don't click the first result for your search, scroll down to a small channel (usually a lyrics one) without many views. These channels are almost never monetized, meaning you can listen to your videos without fear of ads!"

(2) **Scamming return policies/rewards programs:** Another large class of ULPT posts included advice on how to exploit the return policies and rewards programs of various businesses. Examples include:
- "Create a Nike Membership with 12 different emails for 30% off year round."
- "If you need to get a new refrigerator filter, buy a new filter, then put the old filter back in the package and return it saying you bought the wrong one."

(3) **Get out of work:** There is a lot of advice in ULPT on how to get away with doing little to no work at your job. Examples:
- "Working from home and need to appear online? Prop up a lock on the period button within the note pad application."
- "Don't do much at work? Occasionally change your status to "In a Call/Meeting" to keep them thinking you're doing something"

(4) **Interpersonal deception/spite/prank:** This is a broad category. Behavior including deception, eavesdropping, lying, manipulation, emotional abuse is covered by this category. The tips in this category are less for monetary or material gain and more for spite or some sort of social/emotional/intangible advantage over another individual. Some examples include:
- "Long line in security at the airport? Just move past people while loudly repeating "I'm so sorry, I'm so sorry, my plane leaves in 5 minutes, I'm really sorry.""
- "Government ordered social distancing is the best time to check in on those long time "friends" that you always avoid hanging out with."

(5) **Getting something for nothing:** This topic includes advice for how to get something for free or for less than full price/effort. Samples:
   - "If shipping packages through USPS, use the self service checkout and when you weigh your item, lift the corner of your package off the scale for a cheaper shipping rate"
   - "Wanna pay low price for everything? Find a 1$ item in any physical store and take the barcode. Go to self checkout and slio the barcode in front of the item you dont wanna pay much for."

(6) **Car related:** There is also a large representation of advice regarding getting out of traffic/parking violations and similar situations. Examples:
   - "Pass a cop while speeding on the highway and they start turning around to pull you over? Call 911 and report a drunk driver a few miles behind you."
   - "Avoid having to pay a parking ticket by paying it twice"

2.3.3. *COVID-19.* Another interesting feature of the data is the presence of COVID-19-related tips from both subreddits. The LifeProTips subreddit is more active, so the 4945 posts span approximately the month of December 2020 (plus the first few days of 2021). The UnethicalLifeProTips subreddit is less active, and the 4926 posts span approximately February 2020 through New Year's 2021. Below is one example from each subreddit–we will leave it to the human classification engines reading this to determine which quote came from which subreddit.

- "If you want a COVID-19 antibody test free of charge, the American Red Cross provides COVID antibody results from your donated blood."
- "Can't afford to get tested for coronavirus? Cough your lungs out in a public space and check the news in a day or two to see if anyone tested positive."

2.3.4. *Humor.* The nature of posts from UnethicalLifeProTips seems to vary more when compared to those of LifeProTips. The posts in LifeProTips are for the most part sincerely trying to provide some good advice on various aspects of life. On the other hand some posts in UnethicalLifeProTips are ideas for things someone would actually try (like the refund scams or the tips for getting out of work), but there are also unethical posts that are so impractical or transparent that they are meant to entertain–often in a tongue-in-cheek manner. For example:

- "If you don't pay for your graduation photos, you get a version with the word "proof"watermarked over your face; in which case, you also don't have to pay for a diploma."

2.4. **Feature Engineering.** We use `WordNetLemmatizer` together with `word_tokenize` from `nltk` to lemmatize the words from the text data in both the `title` and `selftext` columns.

We next move to vectorize the text data. We first split the data into training (75%) and testing sets (25%) to avoid any data leakage in the vectorization step. The vectorizers are evaluated on and fit to exclusively training data. We consider `sklearn`'s `CountVectorizer` and `TfidfVectorizer` and use a preliminary `MultinomialNB` model (also from `sklearn`) to find the best vectorizer. We look at several different combinations of `title` and `selftext` with and without bigrams to find the strongest way to vectorize the data. We engineer a new column, `alltext` which takes the two strings from `title` and `selftext` and concatenates them into a single string. The best results are obtained when we vectorize `title` and `alltext` individually using `CountVectorizer` and including the top 3000 bigrams. Following a gridsearch, the minimum document frequency we use is 10.

2.5. **Most predictive words.** By fitting a `CountVectorizer` to the `title` training data alone, and predicting on an identity matrix, we are able to obtain, for each word in the corpus, the probability that word is predicted to be part of a post from the UnethicalLifeProTips subreddit. Figure 3 indicates the most predictive words for each subreddit.

The top three words are "resolution," "goal,"and "habit.". Each word in the wordcloud aligns (allowing for seasonal considerations) with the observation that posts from LifeProTips are of a self-improvement nature.

The top three predictive words for UnethicalLifeProTips are "refund," "toilet,"and "fake."The presence of "refund"and "fake"and many others in the wordcloud is not surprising considering the theme of dishonesty underscoring most of the posts in the subreddit. However, "toilet"needs some more explanation: it turns out that there are many posts on how to get free toilet paper. These align with the toilet paper shortage during the early days of the COVID-19 pandemic. For example:

- "Need toilet roll? Public bathrooms, free and fully stocked."

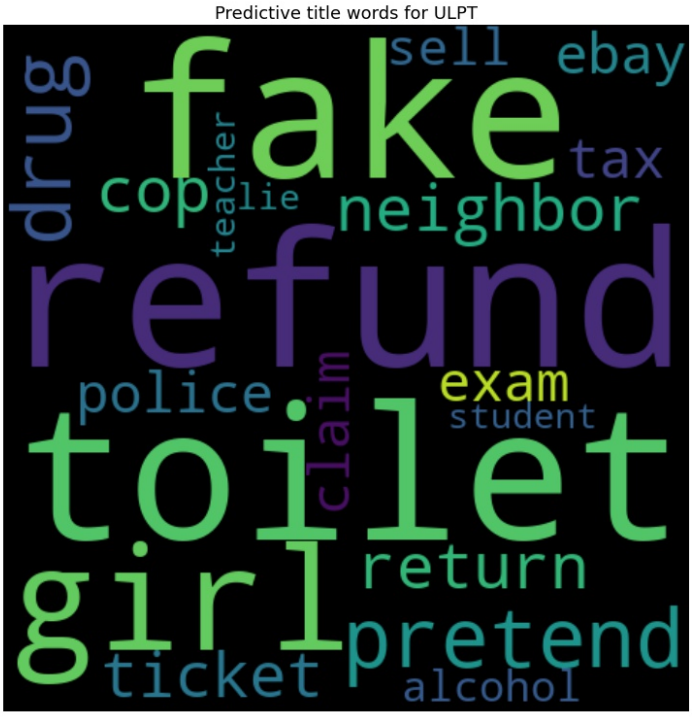

FIGURE 3. Wordcloud for most predictive `title` words for each subreddit

TABLE 2. Cross-validation results

Estimator	Hyperparameters	Mean CV Accuracy	Test Accuracy
`RandomForestClassifier`	`criterion='gini', max_depth=78, max_features='log2', n_estimators=143`	0.7614	0.7880
`MultinomialNB`	`alpha=1`	0.7772	0.7917
`LogisticRegression`	`C=0.1, dual=False, penalty='l2', tol=0.0001`	0.7714	0.8022

TABLE 3. Confusion matrix for `RandomForestClassifier` with threshold 0.5718 predicting ULPT

	Predicted False	Predicted True
Actual False	47.0%	3.2%
Actual True	24.6%	25.2%

Another word's appearance in the wordcloud is slightly puzzling at first: "girl."The word "girl"turns out to be the fourth most predictive word for UnethicalLifeProTips. From reviewing the data there are several different themes for this word on the UnethicalLifeProTips subreddit. One theme is posing as a girl on the internet:

- "Pretending to be a girl online to get free stuff"

Another major theme is "dating"advice:

- "Save your side guy or girl's number in your phone as Potential Spam. This way it will not cause alarm if your boyfriend/girlfriend sees them calling you."
- "See an pretty girl at the grocery but too shy to approach her? Go to the pet food aisle and put the biggest bag of dog food into your cart. Then push your cart down the aisle she's on and boom, she'll probably start talking to you first."

3. MODELING

3.1. **Model assessment.** We first apply the selected vectorizers to the `title` and `alltext` columns in the training data and concatenate using `scipy.sparse.hstack` together with the `title_reading_level` column (passed as a sparse matrix). We then perform 5-fold cross-validation with three different estimators: `RandomForestClassifier, MultinomialNB, LogisticRegression` in `sklearn`; for `MultinomialNB`, we ignore the `title_reading_level` column. We score the cross-validation with both the accuracy metric and the ROC-AUC metric. Since the data is balanced, accuracy is a meaningful metric. Table 2 provides the results from these cross-validations along with the performance on the test data.

3.2. **Threshold tuning.** The next step is to tune the threshold for each estimator since each estimator in truth returns a probability value for its predictions. We need to select the best threshold for each of our use-case. We wish to use this model as an automated filter for unethical content. Such a model would be used by social media moderators, internal business communications, online forums, etc. The automated aspect of the model is intended to significantly reduce manual monitoring, so we choose to prioritize reducing false positives (a positive label belonging to the UnethicalLifeProTips subreddit). This means we place more emphasis on obtaining a higher precision at the cost of a lower recall. After examining the precision-recall curves for the three estimators above, we choose the `RandomForestClassifier` estimator with threshold 0.5718 to obtain a precision of 89.03% with a recall of 50.20%. Figure 4 shows the precision recall curve for the `RandomForestClassifier` estimator, and Table 3. This means that out of all of the posts predicted to be from the UnethicalLifeProTips subreddit, just under 9 out of 10 are truly from the ULPT subreddit, and out of all the posts that are truly from the ULPT subreddit, about 1 out of 2 are flagged by the model.

4. CONCLUSION

For most of the posts in this dataset, any individual with a reasonable amount of cultural and ethical awareness can accurately select which subreddit a given post comes from. The data indicate that a piece of

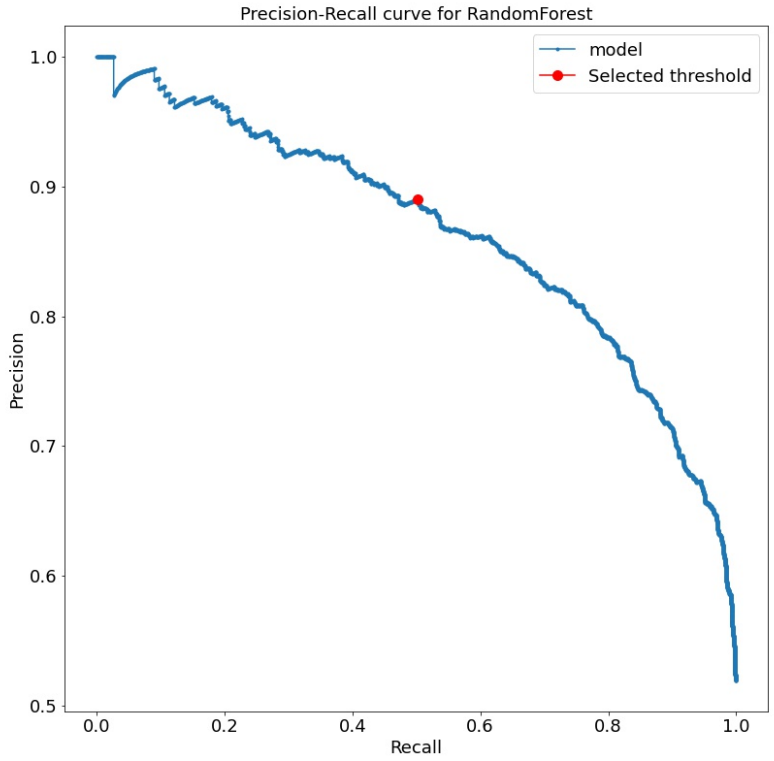

FIGURE 4. Precision-recall curve for `RandomForestClassifier`, predicting membership of a post in the UnethicalLifeProTips subreddit. The position of the selected threshold is indicated in red.

advice offering self-improvement that can be achieved by an individual alone is likely to come from LifeProTips, and a piece of advice suggesting dishonest behavior to gain something from others is likely to come from UnethicalLifeProTips.

Overall LifeProTips reading levels are slightly higher than those of UnethicalLifeProTips. Thanks to our clustering via `LatentDirichletAllocation`, some common themes for these subreddits have been recognized:

- **LifeProTips:** Resolutions, Cooking/kitchen, Money, Gift Giving, Mental health, Cleaning, Online behavior
- **UnethicalLifeProTips:** Avoid paywalls/ads, Scam return policies/rewards programs, Get out of work, Interpersonal spite/prank, Get something for nothing, Car related

The top three predictive words for each subreddit are as follows:

- **LifeProTips:**
  (1) "resolution"
  (2) "goal"
  (3) "habit"
- **UnethicalLifeProTips:**
  (1) "refund"
  (2) "toilet"
  (3) "fake"

We obtained a binary classification model with precision 89.03% and recall 50.20% predicting a post's membership in the UnethicalLifeProTips subreddit. Such high precision indicates that the model is effective in that when it identifies a post as unethical, it is correct 9 out of 10 times. This greatly reduces the need for manual monitoring for forum moderation. The recall, much lower than the precision, shows that 1 out of every 2 unethical posts goes undetected.

It is worth noting that the content from the subreddits is in the form of advice. So this model is trained to specifically recognize unethical advice rather than more general unethical content. Furthermore, the motivation for writing each post seems to differ across the subreddits: LifeProTips are more sincere, UnethicalLifeProTips are a mix of sincerity, spite, and humor.

This model can be strengthened and made more accurate with more data. We could scrape many more posts from the two subreddits to provide more training data and make the model more robust. We could also set up a remote server to periodically scrape more data from the subreddits and automatically retrain the model.