Paul Galatic
CSCI-331
Project 2

Introduction

Subreddits cater to communities of a specific purpose. Given this fact, this report examines the method by which a machine may be taught to tell them apart. Five classification algorithms are tested, and the statistics of their results are graphed. Finally, a word cloud is presented as a visual representation of the contents of the subreddits.

Method

This report primarily relies upon Python 2.7 and the sklearn package.

1. Download Reddit data. The author of this report used the Python Reddit API Wrapper (PRAW) package. Reddit data was downloaded from these two subreddits:
    a. /r/LegalAdvice (~11k posts)
    b. /r/relationships (~12k posts)
2. Split data into Train, Dev, and Test sets. Train and Dev sets were used for training and evaluation during development. Test sets were only utilized for the evaluation of the models for this report.
3. Before training, individual posts are turned into "bags of words" using sklearn's CountVectorizer and TfidfTransformer.
4. These five classifiers were utilized:
    a. Stochastic Gradient Descent Classifier
    b. Support Vector Classifier
    c. Support Vector Regressor
    d. Random Forest
    e. Logistic Regressor
5. The Support Vector Regressor had its results binned, with some examples shown below. It was the only model with a float output.
6. Word maps were created using the Python wordcloud package.

Results

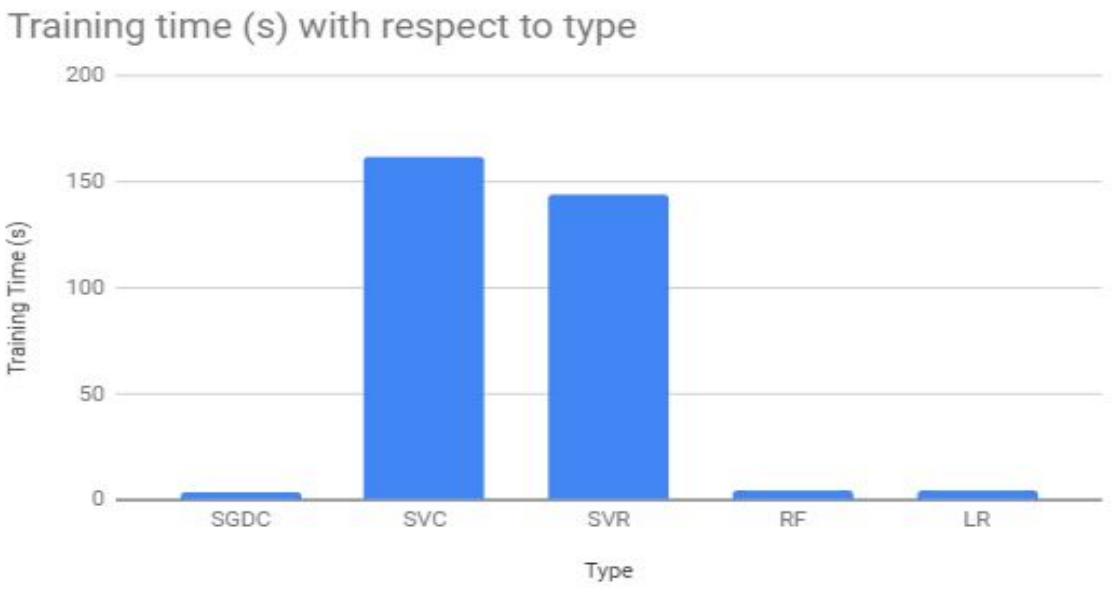

figure 1 - Training time of the models. This is variable, and shown more as an illustration of magnitude rather than as concrete data. The SVC and SVR take much longer to train than the other three.

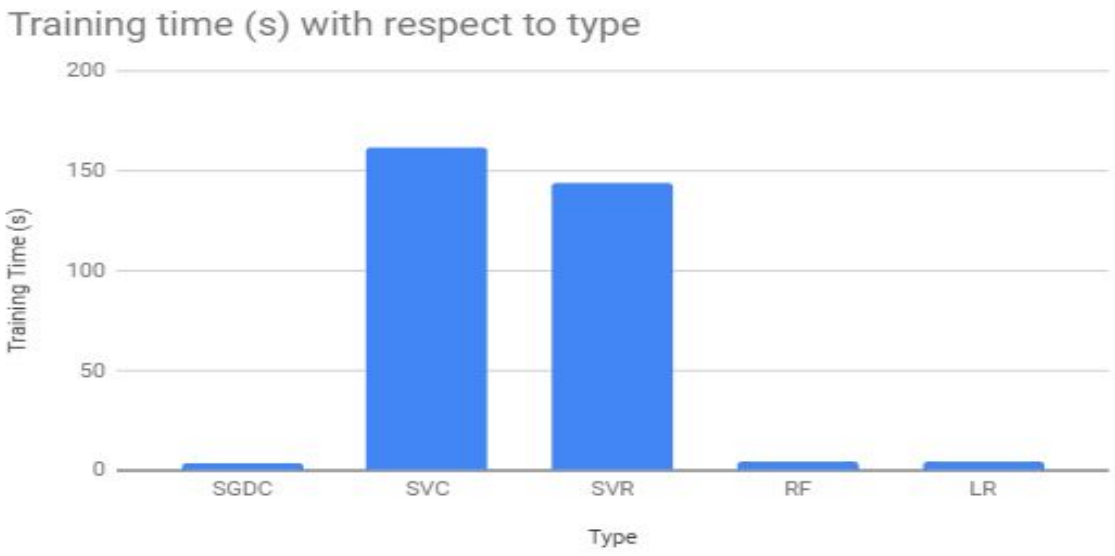

figure 2 - Size of models in kilobytes. As with figure 1, this is more an illustration of magnitude.

figure 3 - Model performance during training. Note the altered minimum value of the vertical axis. The Random Forest performs particularly well during training and somewhat overfits.

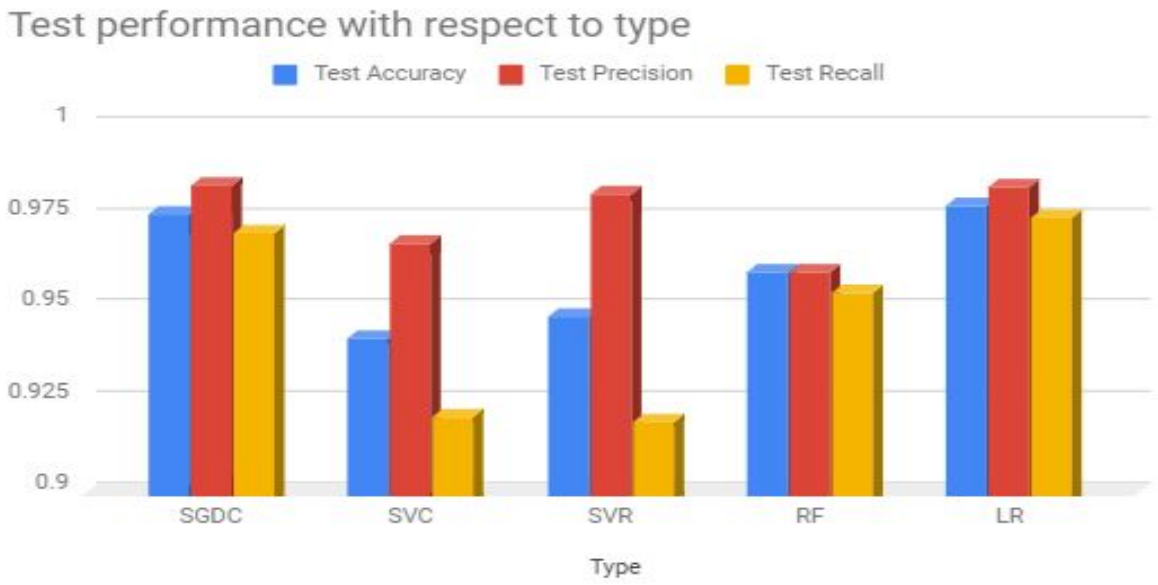

figure 4 - Model performance during testing. Axes are the same as figure 3. SGDC and Logistic Regression perform the best here, though all models perform reasonably well.

One advantage to using the SVR classifier was that it reported results in the form of confidence, rather than a binary label. There may be ways to extract this type of confidence from the other models, particularly logistic regression, but this was the simplest. Below I will share some insights about the distribution of the data and what made certain posts more difficult to identify than others.

figure 5 - A makeshift histogram, as Google Sheets wasn't cooperating. A confidence score of 0 means the SVR believes the post came from /r/LegalAdvice, and 1 means /r/relationships. Items in bin '0' had scores less than '0.1', and those in '0.1' had scores less than '0.2' but greater or equal to '0.1', et cetera. This graph was created by analyzing only the training data.

The distribution of the data is roughly bimodal, which is to be expected from an accurate classifier. This being said, there are many posts that the algorithm is not confident in.

As the posts themselves tend to be very long, excerpts are sometimes used for comparison below.

Bin 0 (Certain /r/LegalAdvice):

Correctly Classified

Brother was charged with OVI in Utah for marijuana (stopped for speeding). They took his rental car, 18k cash and his cell phone. They found a joint in the car. Is taking his cash and phone legal?

Incorrectly Classified

No examples.

Bin 1 (Extremely confident /r/LegalAdvice):

Correctly Classified

(CA) School putting up cameras in the bathrooms, what action can I take I know that this has been posted here before at a school in pal alto, but I’m trying to see if there is any specific law this violates in California that I can site to justify protesting the cameras.

Incorrectly Classified

[28/M] looking to move in with [26/F]'s house. How to Address Rent/Utilities/Etc X-posted in legaladvice and personalfinance Throwaway for obvious reasons.State: MIHi all!I am trying to navigate the intricacies of moving in with my girlfriend, who owns her own house. We have been dating for about a year and a half and have begun talking about moving in together. She recently bought a house and has a monthly mortgage payment of $750/mo (includes taxes and insurance). I make a little more than her (\~5k) but while I have no debt, she has quite a bit of it (around $80k, with most of it being student loans with some credit card debt). Her payments are easily over $1k/mo for her loans.Now, we so far have agreed to split groceries and utilities. What is currently the issue is rent and her definition of "basic upkeep" of the house (repairs/fixes) vs upgrades.With regard to rent, I have asked for the following:1. The rent is split down the middle at the current amount she's paying ($750/mo). I would pay $375.2. If her taxes or insurance increase, I do not bear the brunt of any increases...

Bin 9 (Certain /r/relationships):

Correctly Classified

I broke up with my best friend/girlfriend today (21M) & (20F) Today I decided to end things after a year and 1/2 of dating. My reasoning was due to us being so young and incapable to focus on our goals, we are constantly worried about one another too much. I feel like I am missing out on my youth as well, but it hurts because it ended with us still loving each-other. I know she's crying her eyes out right now, just sucks. I feel so sad, I just want her in my arms right now. I don't know whether I made the right decision, but I feel like this may the best thing for me (possibly is). Only time will tell. Tl;Dr I need some good advice on the best way to get over a person you love.

Incorrectly Classified

No examples.

Bin 8 (Extremely confident /r/relationships):

Correctly Classified

I've [21/f] been seeing this guy [22/m] that has a girlfriend, what should I do? I've been seeing this guy for 4 months now. He has a girlfriend of almost 2 years. It makes me feel bad because I see pictures of them go up online and how see says "I can't wait to see my man." He bought her flowers too. He never mentions me to anyone and that makes me feel bad. I'm not some secret and I refuse to let him think I am...

Incorrectly Classified

[OH] Ex-Girlfriend may be insinuating that towards the end of our relationship that I took advantage of her while she was upset and didn't want to have sex. So I recently broke up with my girlfriend of almost 4 1/2 years. We went through a lot together, and although the relationship's failure was almost entirely my fault, I have a lot of respect for her. We met in high school and for a period of time I was 18 and she was 16 (Two years in, and in case it's relevant)...

Bins 4, 5 (Toss-up):

originally from /r/LegalAdvice:

A friend's xboyfriend won't move out A female friend in Florida ended her relationship in July with her live in boyfriend and he won't move out. She owns the house, he does not pay rent and there is no lease. How can she get him out, she doesn't feel safe in her own home. What to do?

I'm not sure what to do... I work for I.h.s.s taking care of my brother who is disabled and autistic. I love my brother very much and I don't want him to end up going to a nursing home. I dont like it when I feel like people are picking on him or neglecting him. That's why I love and care for him. But I recently caught a case and will be doing time for 45 days and I'm not quite sure what I need to do to make sure that when I come out I will still be caring for him. I have somebody that is willing to care for him while I'm gone. Basically I'm not quit sure what steps I need to do from here... Any advise will help Thx for reading... God bless you

originally from /r/relationships:

My (30F) husband (38M) wants to take a substantial pay cut to help his brother My husband currently has a job making $45,000 which is good paying for the area. Recently my husband's brother has decided to start a new business. He wants my husband to quit his job and work at the new business for roughly $21,000. The main problem is we can't take this kind of hit financially. I told my husband this and he just insisted that he needs to help his brother. This job requires no education and my husband has no previous experience with it. It’s a job basically anyone can do and his brother could easily find another person to fill the role. **tl;dr My husband always wants to be the good guy even if it means throwing himself (and me) under the bus. I feel like his brother is taking advantage of my husband's doormat personality.

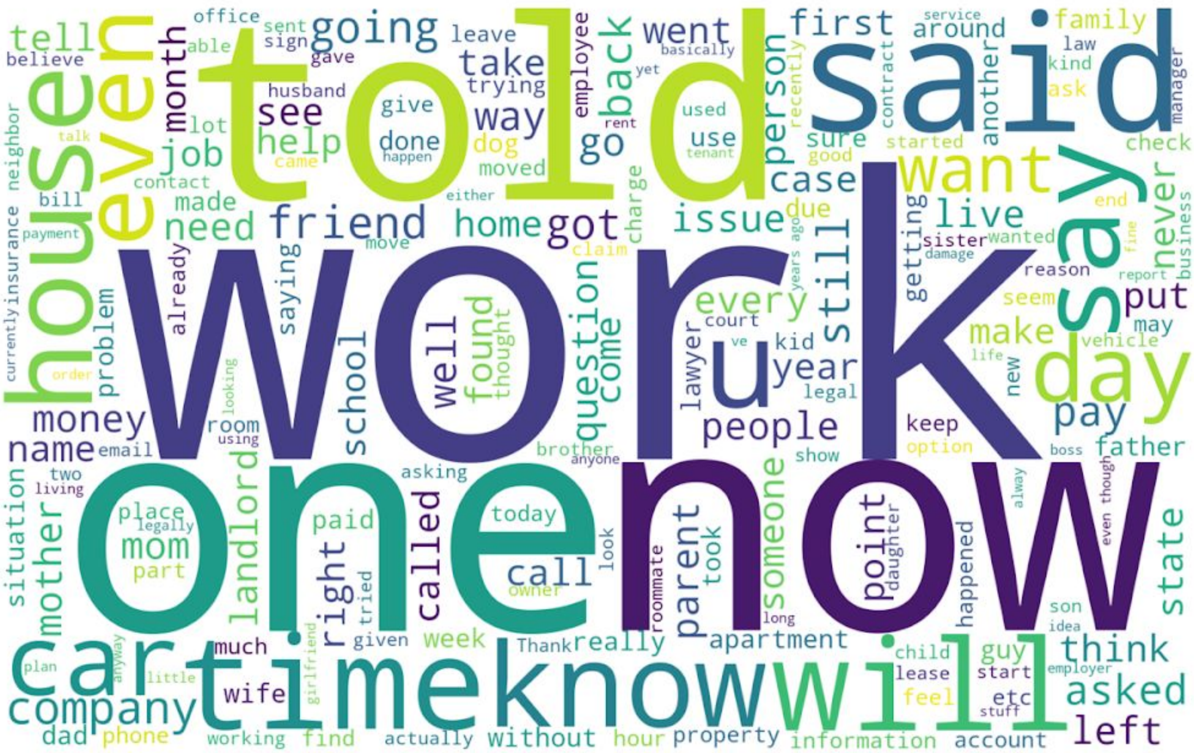

figure 6 -/r/LegalAdvice wordcloud

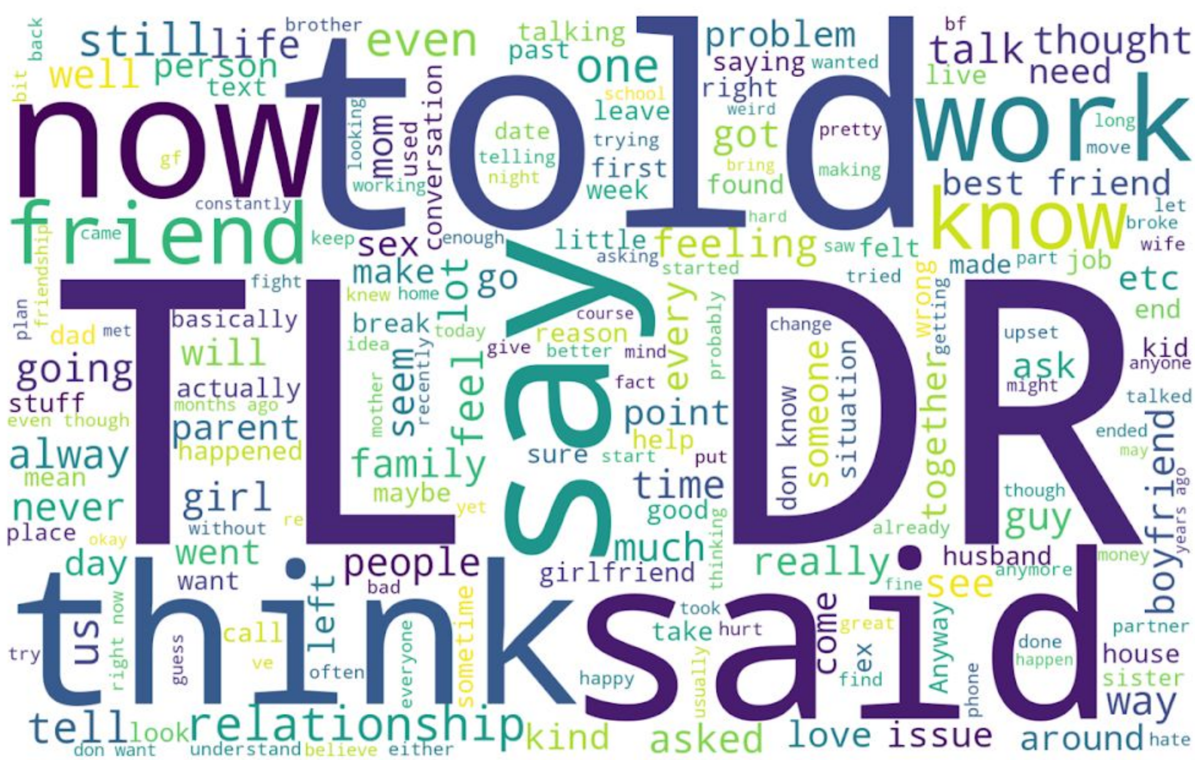

figure 7 - /r/relationships wordcloud

Conclusion

Logistic Regression Classifiers and Stochastic Gradient Descent Classifiers seemed to have the best overall performance with this dataset, in terms of both performance and development time. SVC and SVR classifiers needed to painstakingly consider all of the available data, while the others only take portions in order to minimize their predictive loss (or, in the case of logistic regression, to choose features randomly).

This being said, all models performed reasonably well, as these subreddits demonstrated several notable differences in terms of the topics they focused on.

Posts in /r/LegalAdvice tend to involve money, employment, and local ordinances or politics. Jobs, landlords, local ordinances, and law enforcement are all common topics here.

Posts in /r/relationships are predominantly about family, boyfriends, girlfriends, and other forms of interpersonal connections. They also typically end a post with the phrase "tl;dr" as the posts in /r/relationships can be extremely wordy.

The prevalence of "tl;dr" is illustrated perfectly in the /r/relationships wordcloud above. The wordcloud for /r/LegalAdvice focuses much more heavily on significant, but pedestrian details of living ("house", "car", "company", "work"). /r/relationships also has "work", but focuses more on "relationship", Both subreddits emphasize what others have communicated ("told" and "said"). Surprisingly, in /r/relationships, "friend" seems to occur more frequently than both "boyfriend" and "girlfriend", though that may be a quirk of how the wordcloud was generated.