Web Info Extraction & Text Classification

Antony Paulson Chazhoor

Purpose

“Can a machine be trained to read random text on the internet & identify what it is about?”

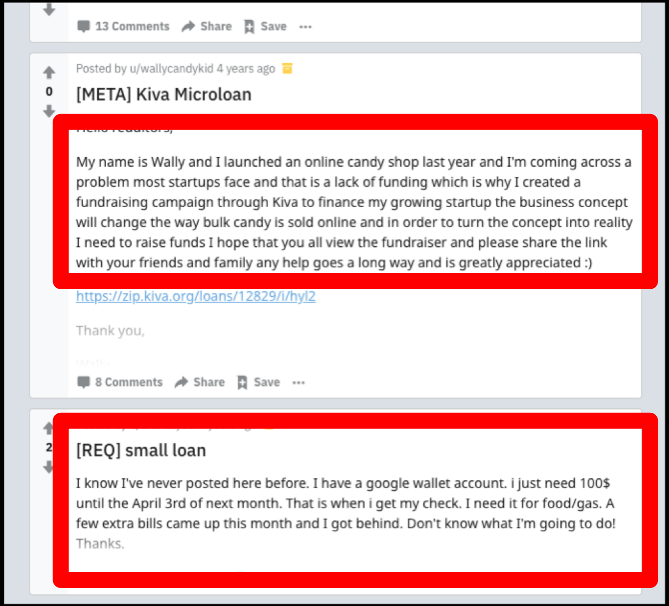
“Reading through entire texts is a cumbersome process. Can machine learning help to make this process easier ?”

“Loans” & “Credit Cards”: I address the questions by choosing two very similar but equally diverse topics within banking

Project Highlights

- Reddit is a popular website which contains blog posts and a lot of information on various topics.
- Two specific subreddit pages scraped
  - Loans: https://www.reddit.com/r/Loans
  - Credit Cards: https://www.reddit.com/r/CreditCards
- Exploratory data analysis on word distributions in both posts
- 3 different machine learning classification models applied
  - Logistic Regression
  - Gaussian Naive Bayes
  - Densely connected Neural Networks
- ✓ Evaluation of models (Classification Accuracy & Confusion Matrices

Scraping of Reddit Posts



Posted by u/wallycandykid 4 years ago

[META] Kiva Microloan

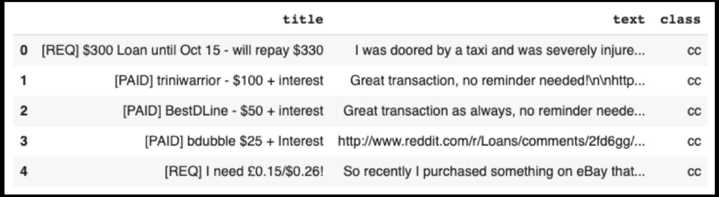
My name is Wally and I launched an online candy shop last year and I'm coming across a problem most startups face and that is a lack of funding which is why I created a fundraising campaign through Kiva to finance my growing startup the business concept will change the way bulk candy is sold online and in order to turn the concept into reality I need to raise funds I hope that you all view the fundraiser and please share the link with your friends and family any help goes a long way and is greatly appreciated :)

https://zip.kiva.org/loans/12829/i/hyl2

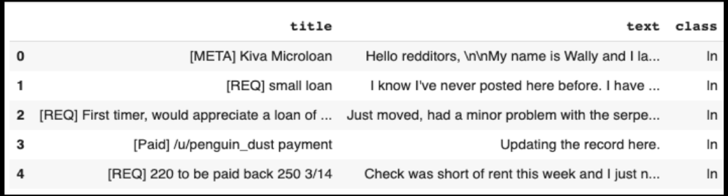
Thank you,

[REQ] small loan

I know I've never posted here before. I have a google wallet account. i just need 100$ until the April 3rd of next month. That is when i get my check. I need it for food/gas. A few extra bills came up this month and I got behind. Don't know what I'm going to do! Thanks.



	title	text	class
0	[REQ] $300 Loan until Oct 15 - will repay $330	I was doored by a taxi and was severely injure...	cc
1	[PAID] triniwarrior - $100 + interest	Great transaction, no reminder needed!\n\nhttp...	cc
2	[PAID] BestDLine - $50 + interest	Great transaction as always, no reminder neede...	cc
3	[PAID] bdubble $25 + Interest	http://www.reddit.com/r/Loans/comments/2fd6gg/...	cc
4	[REQ] I need £0.15/$0.26!	So recently I purchased something on eBay that...	cc



	title	text	class
0	[META] Kiva Microloan	Hello redditors, \n\nMy name is Wally and I la...	ln
1	[REQ] small loan	I know I've never posted here before. I have ...	ln
2	[REQ] First timer, would appreciate a loan of ...	Just moved, had a minor problem with the serpe...	ln
3	[Paid] /u/penguin_dust payment	Updating the record here.	ln
4	[REQ] 220 to be paid back 250 3/14	Check was short of rent this week and I just n...	ln

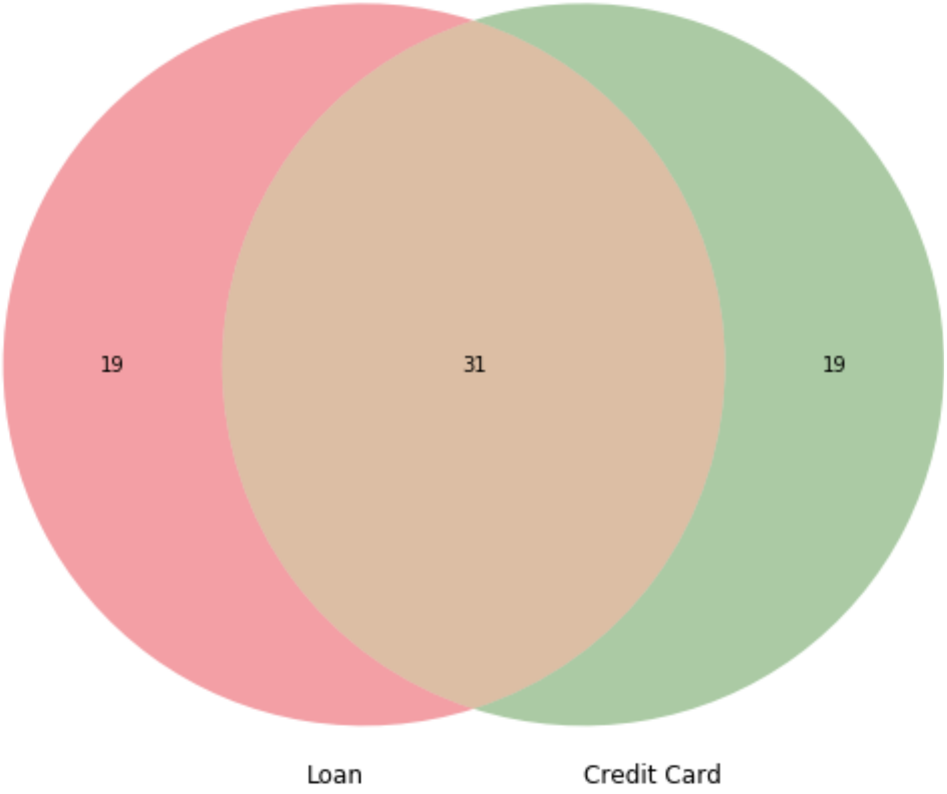
- Each page contains 25 posts
- Scraping done on nearly 50 pages to collect at least 1000 posts

Exploratory Data Analysis

- Checking the word distributions among posts

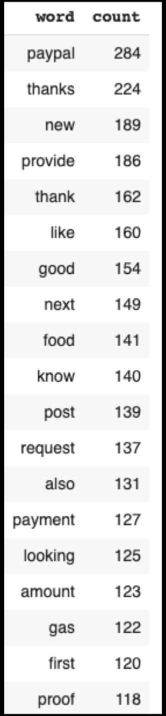
- An intuition into how similar/different they are

- Among top 50 words in both posts 31 were the same.
- There were 19 unique words

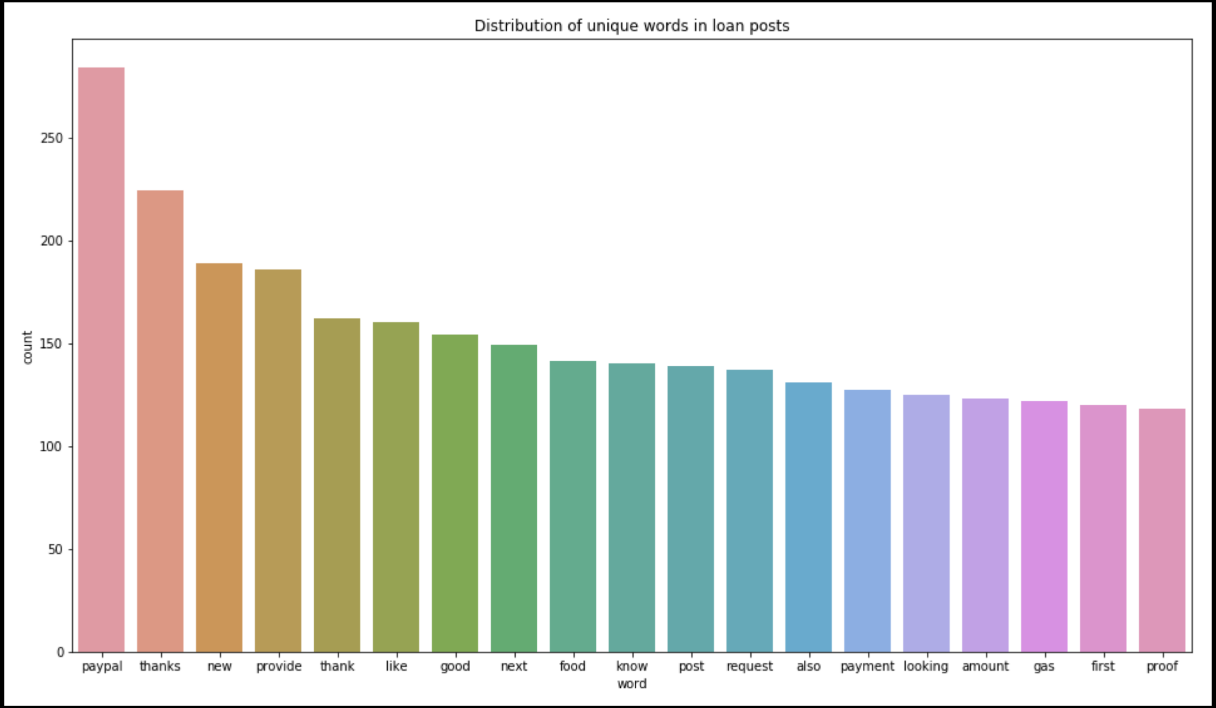


EDA (contd…)

- Top unique words in Loan posts.

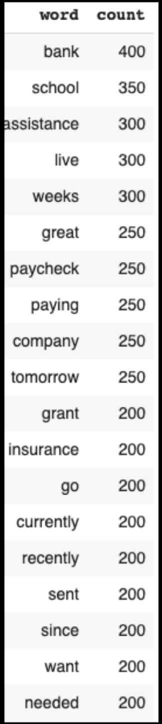


word	count
paypal	284
thanks	224
new	189
provide	186
thank	162
like	160
good	154
next	149
food	141
know	140
post	139
request	137
also	131
payment	127
looking	125
amount	123
gas	122
first	120
proof	118

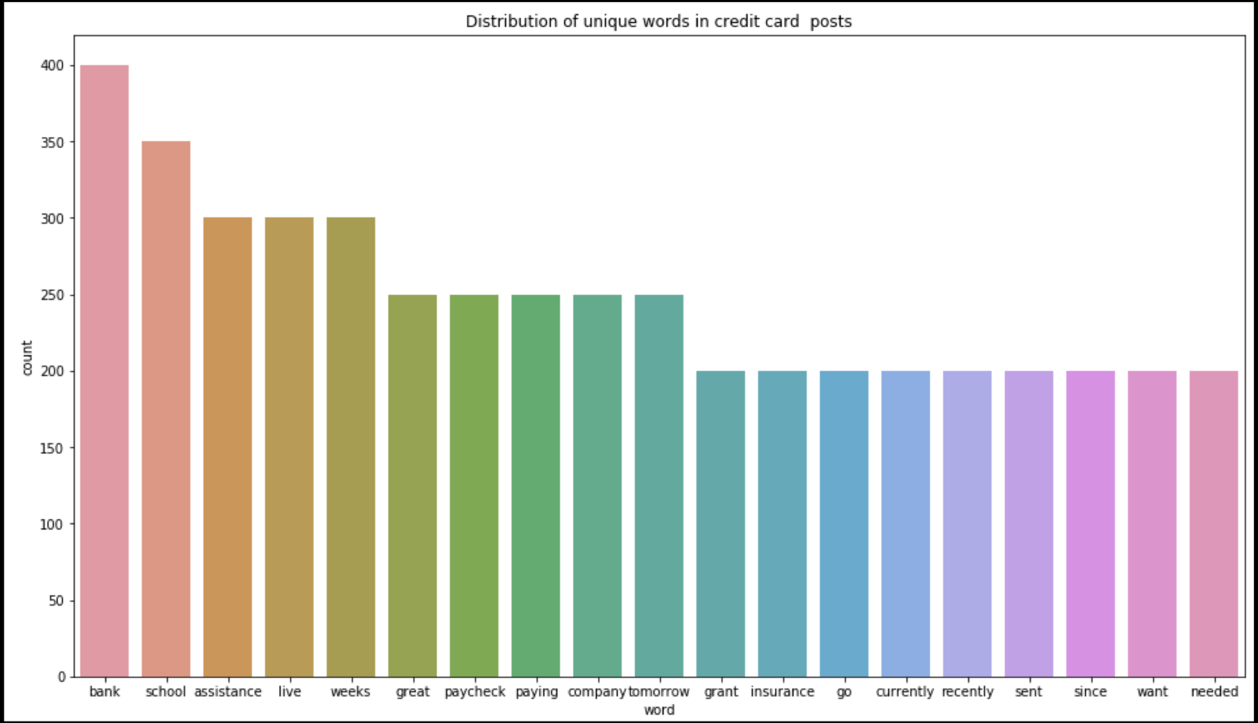


EDA (contd…)

- Top unique words in credit card posts.

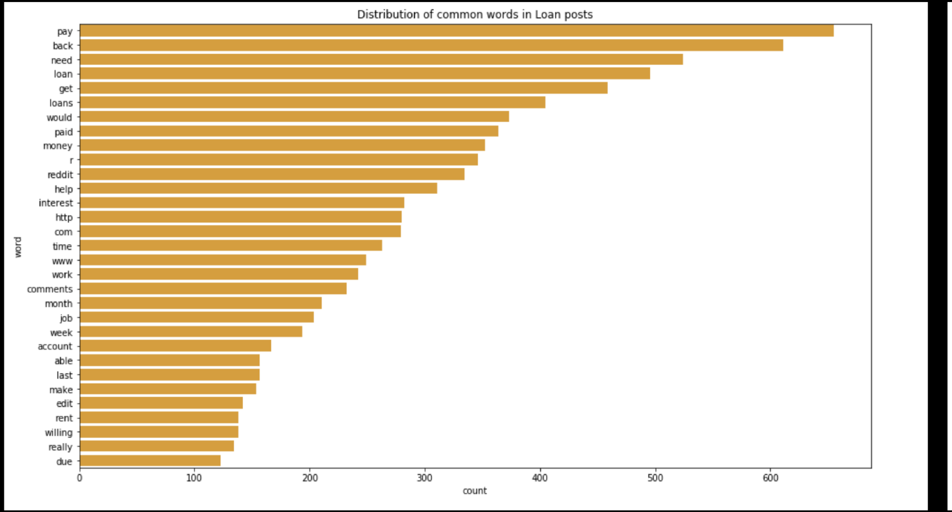


word	count
bank	400
school	350
assistance	300
live	300
weeks	300
great	250
paycheck	250
paying	250
company	250
tomorrow	250
grant	200
insurance	200
go	200
currently	200
recently	200
sent	200
since	200
want	200
needed	200

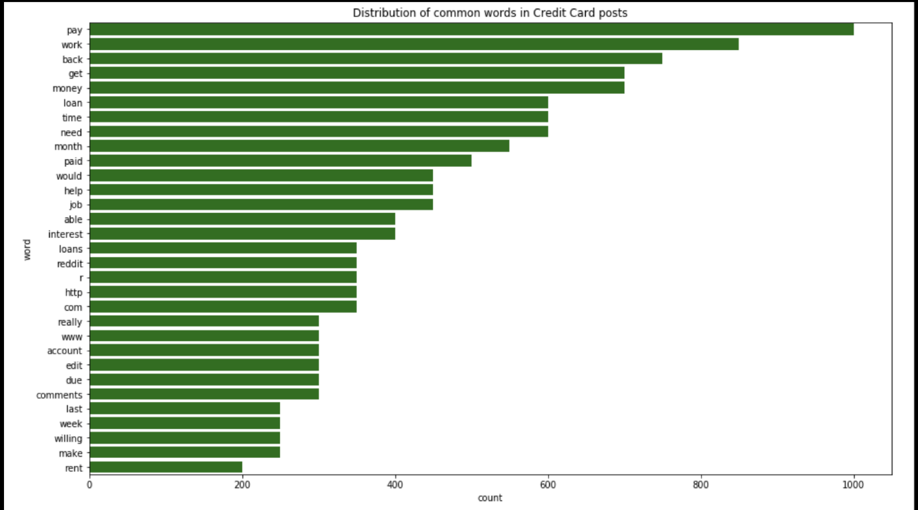


EDA (contd…)

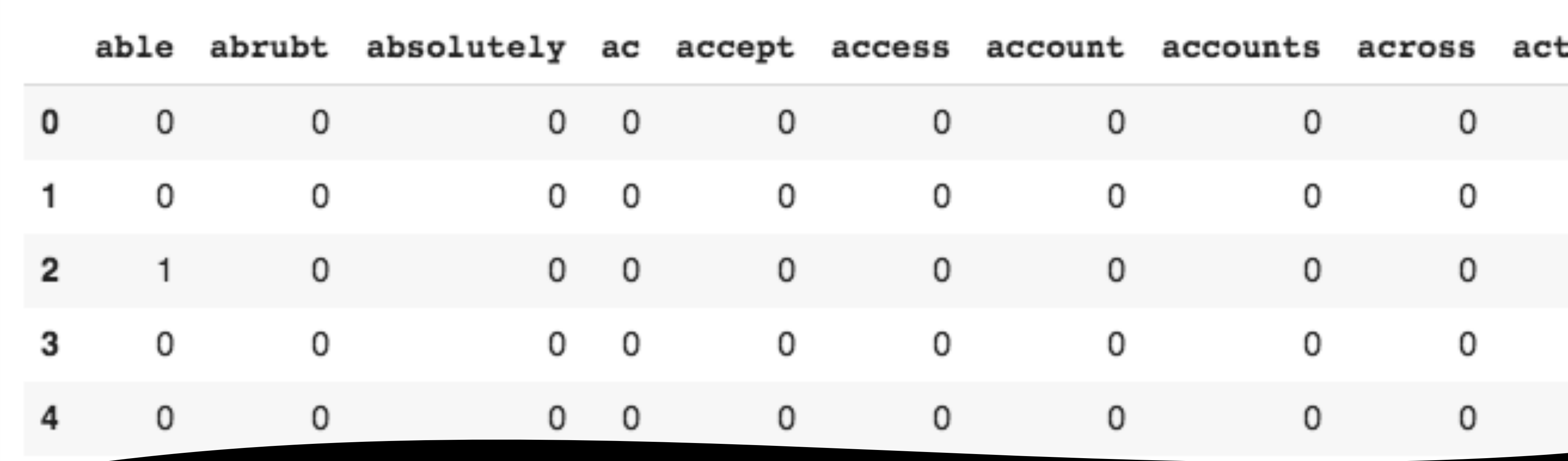
- Common words and their occurence in both posts



Loan Posts



Credit card Posts



	able	abrubt	absolutely	ac	accept	access	account	accounts	across	act
0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	
2	1	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	

Preprocessing

- Texts by themselves can not effectively be added in as features to a statistical model
- The count vectorizer is used to transform this text into a collection of most occurring words and turn them into columns of a model.
- The values for these columns is the frequency of that word in the text

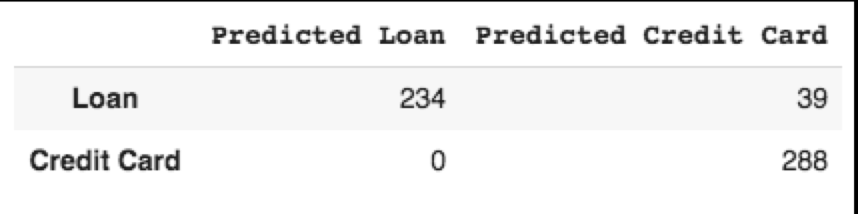
Model Evaluation

- Classification accuracy on the testing set is the parameter by which the models were judged
- With 800 features from the Vectorizer, the models performed exceptionally well
- Logistic regression outperformed the other two models by a narrow margin
- The values for these results of each model:

Classification Model	**Training Accuracy**	**Testing Accuracy**
Naive Bayes	91.91 %	93.04 %
Neural Networks	94.65 %	93.22 %
Logistic Regression (with GridSearch)	95.19 %	95.54 %

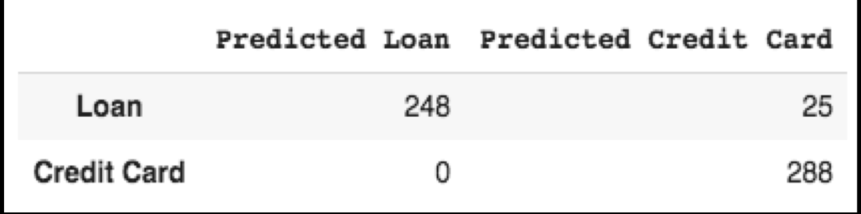
Model Evaluation (Confusion Matrix)

Gaussian Naive Bayes model:



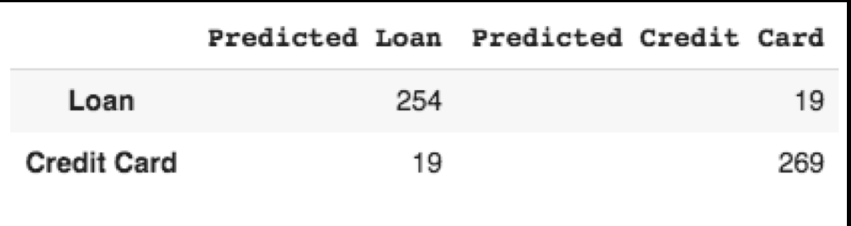
	Predicted Loan	Predicted Credit Card
Loan	234	39
Credit Card	0	288

Logistic Regression



	Predicted Loan	Predicted Credit Card
Loan	248	25
Credit Card	0	288

Neural Networks:



	Predicted Loan	Predicted Credit Card
Loan	254	19
Credit Card	19	269

- The confusion matrices give a much clearer picture of model performance
- Though it misclassified 25 Loan posts logistic regression performed extremely well in identifying Credit card posts

Conclusions and further steps

The project indicates how machine learning can aid in text classification.

These models can save a lot of time spent otherwise in manual reading.

Similar banking subjects like loans and credit cards can be differentiated by machine learning.

A step forward would be to gather posts from more topics and train models to perform multi-class classification.