CLASSIFYING CANADIAN POLITICAL DISCUSSION

GA-DSI PROJECT 3

MOTIVATION

- Reddit is the most active English language internet discussion board (by Alexa ranking).
- Anyone can create a new subreddit in about three clicks.
- Popular communities often schism based on ideological differences or as a reaction to perceived inauthenticity.
- Communities then exist in parallel, both observing the same topic but maintaining different cultures.
- Can we use NLP to train a binary classifier to distinguish between some subreddit and its ideological spinoff?

EXAMPLES

Original	Spinoff
r/OffMyChest (2.2m members)	r/TrueOffMyChest (2.2m members)
r/PublicFreakout (2.8m members)	r/ActualPublicFreakout (448k members)
r/Gaming (27m members)	r/TrueGaming (1m members)

OUR FOCUS: CANADIAN ISSUES

Original: r/Canada	Spinoff: r/OnGuardForThee
739,000 members	110,000 members
3,800 active	1000 active
Created 2008	Created 2017
First popular Canadian discussion subreddit	Originated due to disputes over allegations of racist content and moderation in r/Canada

DATA ACQUISITION

- Comments were used for modelling rather than post titles (post titles were often just news headlines).
- One hundred comments per week sampled from each subreddit from Sept 2018 to Sept 2020.
- Moderation bot messages, [deleted], and [removed] comments were purged.
  - 20,102 comments
  - 50.4% r/OnGuardForThee
  - 49.6% r/Canada

PREPROCESSING

- 75-25 train test split.
- Custom stopword set made that included most common words shared between the subreddits.
- Emphasis was on tokenizing words that were unique/had usage skewed to one subreddit over the other.
- Snowbell stemmer used for tokenization (also known as Porter2)

MODELLING

- Why limit ourselves to just three models?
- All permutations of the following vectorizers and classifiers were grid searched and cross validated.
- Vectorizer:
  - CountVectorizer
  - TfidfVectorizer
- Classifier:
  - Logistic Regression
  - K Nearest Neighbors
  - Naïve Bayes
  - Random Forest/ExtraTrees
  - AdaBoost

This took a while.

NOTES FROM MODEL EXPERIMENTATION

- All models except K-Nearest Neighbors overfit.
- However, K-Nearest Neighbors underperformed most models in accuracy.
- Most model's accuracy between 59-64%.
- GridSearchCV skews to large choices of max_features if left to its own devices: will overfit for marginal increases in accuracy.
- No model significantly outperformed any of the others.

- I hope I'm not just measuring randomness

FINAL MODEL: BERNOULLI NAÏVE BAYES

- 4000 features
- Snowball Stemmer
- Stop words removed

Naïve Bayes methods had better runtime than most of the models.

Multinomial Bayes offered slightly better accuracy and F1, but was more prone to overfitting.

Gaussian Bayes with the TFIDF vectorizer performed significantly worse.

FINAL MODEL: METRICS

	All comments	Long Comments (>20 words)	Short Comments
Total Comments	5016	2822	2204
Actual r/OnGuardForThee	2533	1456	1127
Actual r/Canada	2493	1366	1077
Predicted r/OnGuardForThee	2018	1443	575
Predicted r/Canada	3008	1337	1629
Accuracy Score	0.6146	0.6389	0.5835
Recall Score	0.5160	0.6456	0.3708
Specificity Score	0.7148	0.6318	0.8154
Precision Score	0.6477	0.6514	0.6383
Precision (Canada) Score	0.5924	0.6258	0.5641
F1 Score	0.5744	0.6485	0.4443

THOUGHTS

This model is very skewed when classifying short comments, but performs adequately on comments of length greater than twenty words.

Note: Training the model on only longer comments did not improve performance.

TOKEN PROBABILITIES

Concepts/keywords that were strongly predictive:

r/Canada

- China/Canada relations (Huawei, Taiwan, Meng Wanzhou)
- Taxes/economic keywords (Bond, monopoly, GDP, money, buy, tax, cost…)
- Real estate (Realtor, condo, house, home)
- Impaired driving discussion

r/OnGuardForThee

- Reddit meta-commentary (metacanada, onguardforthee, mod, subreddit, brigade)
- Conservative media personality names (Gavin McInnes, Ezra Levant, Jordan Peterson)
- Leftist loaded language (denier, bigot, misogynist, shill, rag, redneck, incel)
- Tokens directly related to politics (socialist, conservative, vote, elect, left, liberal, party…)

THE MOST R/CANADA COMMENT

Exactly one comment was classified to be in r/Canada with >99.9999% probability.

Serious answer?

1. Huawei is the only vendor that is currently able to roll out 5G tech on a commercial scale and at a much cheaper price than its competitors. You can blame tech stealing or whatever, but those are the facts. Banning Huawei would essentially give the rest of the world a head start in the 5G game.

2. Huawei is already being used extensively in 4G networks by Telus and Bell. So if we ban Huawei entirely, we would have to rip out a lot of the current mobile infrastructure as well.

3. There is currently no evidence of any backdoors to Huawei equipment. Of course this doesn't mean that they can't do it later on, but everything so far including the Nortel case and the microchip case by bloomberg have basically all been based on hearsay.

4. Huawei has multi-million dollar research agreements with UT and UBC.

Or if you're r/Canada: 1. Trudeau has been bought by the Chinese.

What are the security risks:

1. Huawei has a history of cyber-industrial espionage. I personally *believe* the Nortel case to be true because from my perspective it would be stupid for Huawei not to have donen it. When Nortel went under, a lot of engineers and techies would have been looking for new sources of income. After all, they just lost their jobs. It wouldn't have been very difficult for Huawei to simply hire them afterwards and then who knows what information they shared. That doesn't mean there is "proof". I've seen a lot of the several year old Nortel article being spread around, but everybody seems to ignore the fact that it literally says that there is no proof right in the middle of the article. I doubt anybody even read it.

2. China has a law that says Chinese companies must share security information with the government if requested. So legally speaking, if China hypothetically asked Huawei to provide information, they would have to.

3. Huawei users will be dependent on Chinese equipment. We already are, but even more so.

4. According to the Snowden leaks, the US hacked Huawei servers and intended on using Huawei hardware to spy on targets such as Iran, Afghanistan, Pakistan, Kenya, Cuba. Since Huawei is being banned in the Five Eyes, this might indicate that recent US operations have not been so successful, but you can't rule out that the US isn't spying on you with Huawei products either. So you'll likely have both China and the US monitoring you at the same time.

EDIT: I would like to say that I don't disagree that Huawei is a security risk. This is simply pointing out why Canada would hesitate to just say "hey we don't want you hear anymore, get lost" on a moment's notice.

THE MOST R/ONGUARDFORTHEE COMMENTS

31 comments were classified to be in r/OnGuardForTheewith >99.9999% probability. Here's two of them.

What a fucking disgrace. Ford is just going to parrot whatever the Fascist Mango does, isn't he?

That they ACTUALLY stood up and called gender identity a "liberal ideology" shows just how deeply the language of the alt-right is leeching into our politics. These clowns sound like a /r/metacanada user posting a hot take to rationalize their bigotry while they wait for their mom to bring fresh chicken tendies to the basement, not functional adults who have the capacity to lead a province.

This is a very unpopular opinion, especially amongst progressives, but in my view, **we need to unify the left**.

The days of socially liberal(ish) conservatives like Kim Campbell, Mulroney etc are **long, long gone**. We now have shitstains like Bobandy in Alberta and Sheerer kissing Faith Goldy's rotten ass. The literal headof the conservatives was making nice nice with a fucking white supremacist.

So, we could talk about how we don't want a two party system like the US. We could talk about how we all have choices. We could make a million fucking reasons about why we should have two or three fucking center / center-left parties. But in the end, if we keep that up, we will fuck ourselves when Sheerer gets 8 or 12 years of running this country.

Alberta only elected the NDP **because the right was fractured**. They figured that shit out real quick didn't they?

Politics is often about choosing the lesser evil, and right now we can either suck it up or get literal fascists running our country like the US.

CONCLUSIONS

Categorizing comments using a “bag-of-words” approach is hard to do accurately without unique keywords in each category to latch onto.

However, our model performed consistently above the baseline.
Accuracy: 0.615. F1-score 0.574.

Classification strength was maximized when restricted to testing comments longer than 20 words.
Accuracy: 0.639. F1-score 0.649.

Despite the model’s lack of accuracy, I think it can still prove useful as a heuristic.