COMP90049 Project 2: Twitter Trolls and the Tweeters who Love them

I. INTRODUCTION

As Twitter is effective in spreading news and affecting people thought, recently this benefit has been twisted for political propaganda. The latest case revealed was about millions of tweets concerning United States 2016 presidential election were generated by Russian trolls. Linvill et al. (2018) report that certain keywords in tweets associated with trolls behaviour. Hence, this research aim is to further evaluate how tweet text can be used to predict trolls. It presents an extensive comparison on how different treatments on data affect the prediction performance. After a sequential trials, it is known that detect trolls based on their tweet text is feasible with constrains regarding the choice of attributes and troll classes.

II. RELATED WORK

There are numerous researches discussing troll detection. Tayade et. al (2017) identify troll based on sentiment analysis by comparing Naive Bayes, Support Vector Machine (SVM), and Maximum Entropy Classifiers. They highlight that it is crucial to select best features and apply the best algorithm that return high accuracy. They found these methods are promising and return high accuracy. Another research is from Atkinson (2018). He classify some keywords into some identities and find their relation with trolls behaviours along with time the tweets posted, and retweet counts. Linvill and Warren (2018) distinguish the behaviours of five type of trolls including Left Troll and Right Troll according to certain hashtag such as #blacklivesmatter and #maga.

III. METHODOLOGY

This paper use a sample dataset which originally contains three million tweets from 2,848 users contain 223,000 attributes[1]. The raw tweet text then pre-processed by considering English characters and presented as a vector space model. Each tweet is classified as Left Troll, Right Troll and Other. The main question is whether this data useful to predict trolls.

This research evaluates how well tweet words contribute to the prediction using supervised methods such as Naive Bayes, Multinomial Naive Bayes (MNB), Decision Tree, and Random Forest. Naive Bayes and MNB assume independent relationship between each attribute while decision tree allows for interaction between a subset of attributes to the label. Random forest is an extended version of Decision Tree which is expected to return higher accuracy. Tweet text is said to be helpful in identifying trolls if a predictive model based on this data returns high accuracy, even precision and recall, and good interpretability.

All models are evaluated with a holdout approach, hence, datasets are divided into training and development. Data exploration and prediction are built under training dataset. All models will be evaluated using development dataset. The most proper model then applied to the test which does not have label. The proportion of classes as follows:

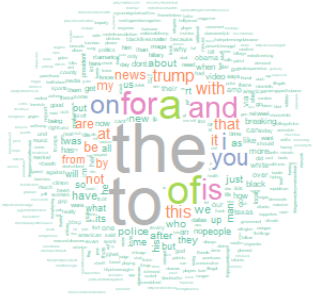
	Training		Development	
	User	Tweet	User	Tweet
Total	100	122,637	42	56,194
Left Troll	16	34,971	7	17,379
Right Troll	29	38,998	15	18,706
Other	55	48,668	20	20,109

IV. RESULTS

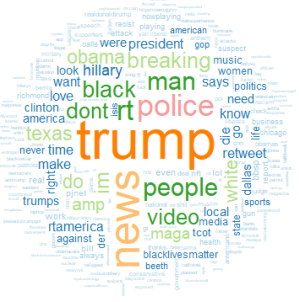
3.1 Attributes Exploration

According to the following word clouds, the stopwords such as *"the"* and *"to"* made the top frequencies. Tayade et. al (2017) remove all of these and focus more on the meaningful terms.

All Attributes

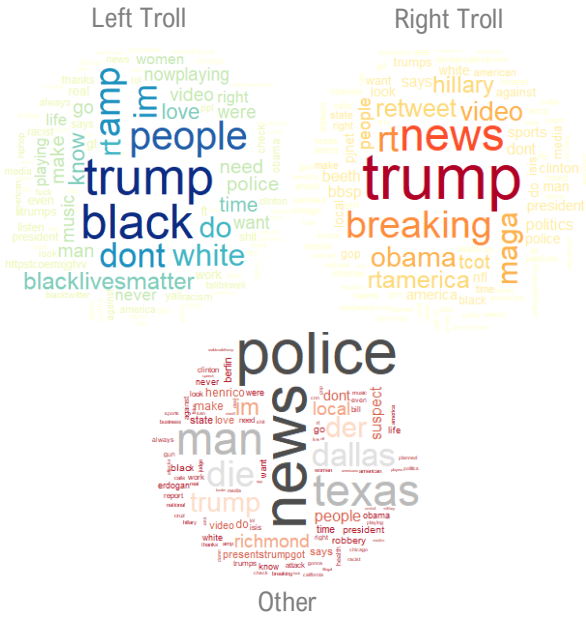


Filtered Attributes

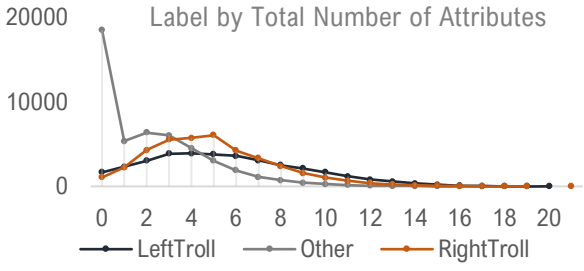


[1] FiveThirtyEight's GitHub

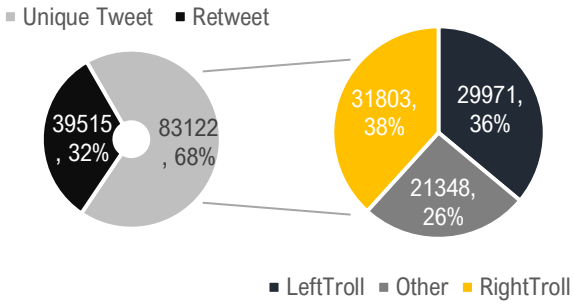
The following word clouds show that certain terms correspond to one class, for example *"black"* for Left Troll, *"breaking"* for Right Troll, and *"news"* for Other. Atkinson (2018) finds that *"black"* term highly correspond to Donald Trump supporter.

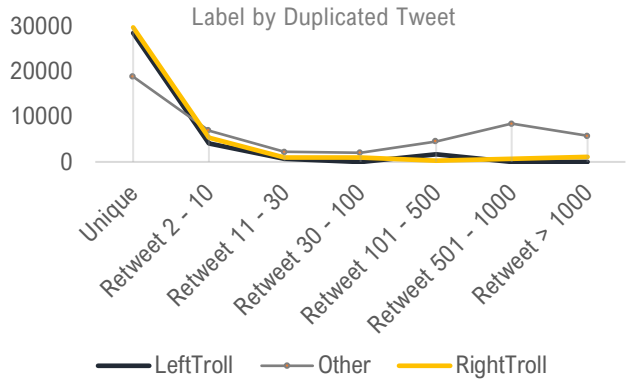


This word clouds illustrate that a great number of words appear nearly even in all classes or have no significant frequency. These are subject to feature selection. However, considering some instances have zero values for all attributes, remove any attributes will account for more instances lie in this manner which may cost the classification accuracy.



Another important aspect to explore is retweet behaviour. Atkinson (2018) discovers that trolls utilize retweet to elevate the likelihood of their tweet being seen. In the dataset, only two-third of tweet text are unique. Both trolls share similar trend in retweet counts which indicates if retweet significantly effective, model built with this manner will difficult to distinguish Left and Right Troll but more into trolls or not trolls. Excessive number of retweet by Other could make the predictions skew to predict this class.





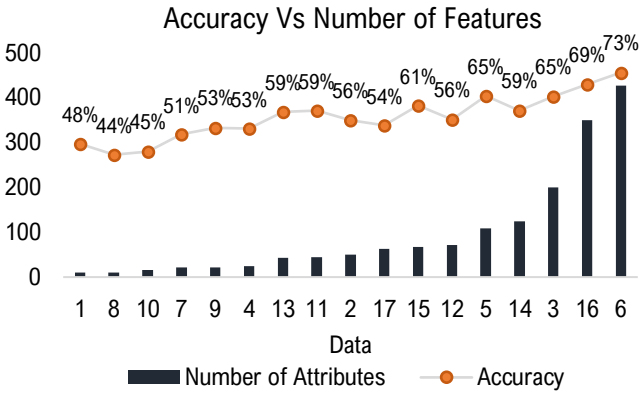
3.2 Predictive Modelling for Classification

The following table shows list of datasets that is derived with different ways of feature selection to see the behaviour of prediction relative to attribute selection approaches. Correlation-based Feature Selection (CFS) only takes those have high correlation with the class and low intercorrelation with other attributes (Hall and Holmes 2003). Mutual Information is similar but not account the intercorrelation.

Data	# of Features	Feature Selection Approaches
1	10	Document Frequency on Original data
2	50	
3	200	
4	25	Mutual Information and Chi-Square on Original data
5	109	
6	427	
7	21	Forward CFS on Data 2
8	10	Stopwords removal on Data 2
9	22	Forward CFS on Data 4
10	16	Stopwords removal on Data 4
11	44	Forward CFS on Data 5
12	72	Stopwords removal on Data 5
13	43	Forward CFS on Data 3
14	125	Stopwords removal on Data 3
15	67	Forward CFS on Data 6
16	350	Stopwords removal on Data 6
17	63	Forward CFS on Data 16

Without considering any attributes Zero-R predict all instances as Other and returns 35.78% accuracy. By using One-R the accuracy slightly increases to 43.95% according to term *"the"* with the highest precision and recall (around 70%) are given to class Other. This become the baseline for the next modelling trials.

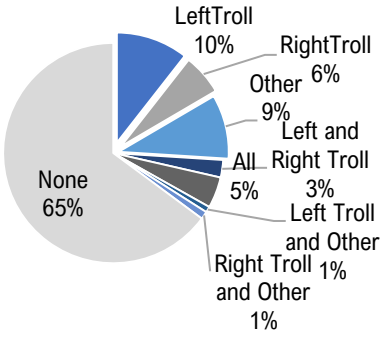
McCallum and Nigam (1998) explain that MNB can reduce error in a considerable proportion for a larger vocabulary size because it accounts the frequency of each term over all terms in each instance, which is match with this case. Both Naive Bayes and MNB hold the "naive" assumption but MNB captures more information. Using the largest dataset, Naive Bayes returns 62.69% which is significantly higher than the baseline but MNB returns much better accuracy with 73.08%. Thus, the next predictive models are built with MNB.



The above chart shows MNB accuracy ordered by total number of attributes. It is noticeable that using more attributes tend to produce higher accuracy but the relative growth is not balance. The smallest dataset (Data 1) contains ten stopwords, such as *"a"*, *"and"*, *"for"*, etc, while the largest dataset (Data 6) contains more stopwords and keywords. Comparing both results implies that adding 4000% more attributes only account for 50% higher accuracy.

This means many of the attributes added not contribute to the model. However, if the attributes are re-filtered, either by remove stopwords[2] or by CFS, the accuracy is dropped. The underlying idea is none of these feature selections works well in returning minimum features with maximum accuracy.

Beside many zero values, this may due to the interdependencies between words that is not accounted in the filter selection, which according to Rish (2001) in some cases it improve Naive Bayes performance. One way to find attribute dependencies is by looking at the MNB predictions for each attributes correspond to each class.



MNB on Data 16 shows 65% attributes not account for any class and other 35% associated with either class, examples as follows. The ones in bold match Linvill and Warren (2018) findings. Some terms appear in pairs, such as names, titles, and race.

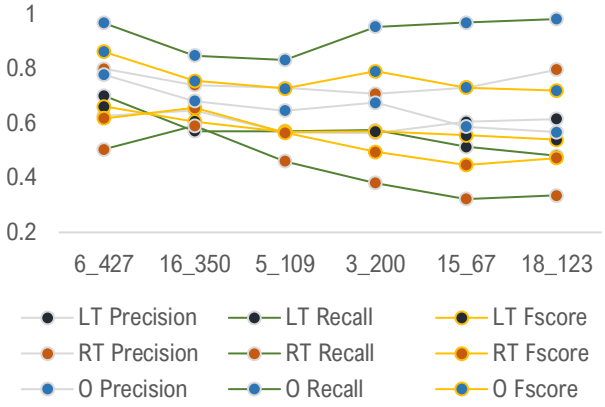
Class	Essential Terms
LeftTroll	**"blacklivesmatter"**, "racist", "women"
RightTroll	**"maga"**, **"tcot"**, "retweet", "rtamerica", "america", "american", *"hillary"*, *"clinton"*, "isis", "politics"
Other	"berlin", "richmond", "chesterfield", "deutschland"
Left and Right Troll	"amp", *"black"*, *"white"*, "right", "rt", *"president"*, *"obama"*, *"trumps"*
All	"people", "police", "time", "trump"
Left and Other	"good", "im", "love"
Right and Other	"local", "news", "state", "texas"

If the 65% are removed (be Data 18) the accuracy is 61.16%. If the attributes correlated with more than one class taken, the accuracy is dropped to 52.5%. This is due to some important terms belong to both trolls are useful for the classification.

Looking at accuracy only is not enough. The graph below visualises performance of the top five models. It clearly seen that Recall of class Other outnumber in all models. Overall, the recall for each Troll is lower than the precision while the recall of class Other is higher than its precision. If an automatic

[2] Collected from https://gist.github.com/sebleier/554280

system built based on any of those model there will be more Trolls misclassified as Other rather than Other misclassified as Trolls which is the cost need to be considered when implement the model.

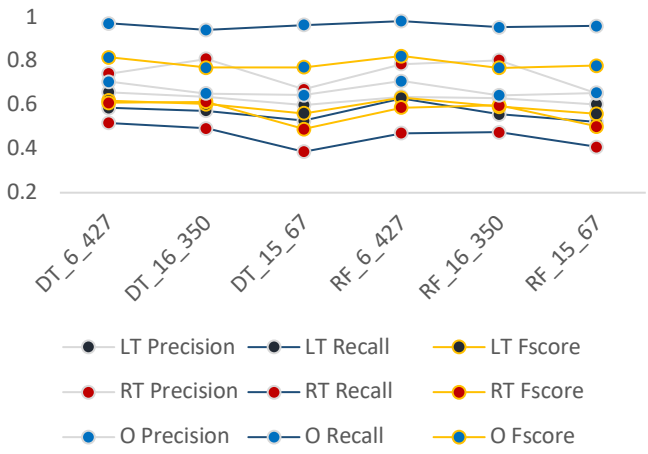


Nevertheless, MNB on Data 16 shows lower but more even distribution of precision and recall. Thus, removing stopwords works better in balancing the precision and recall of all classes. But this does not apply on Data 18 and Data 15. MNB returns lower accuracy and F-scores for data with more meaningful attributes that are highly correlated with each class. This condition suit Rich (2001) finding in which for certain case Naive Bayes performance does not necessarily determined by feature dependencies but more by the cost of independence assumption. Accordingly, embedded method such as decision tree and random forest which account the interactions between attributes and may reduce attribute dimensions are assumed to returns better result.

The following results are from Decision Tree and Random Forest, built under numeric attributes and multiway split according to information gain.

Data	Number of Attribute	Naive Bayes	Decision Tree	Random Forest
6	427	73.08%	70.15%	70.33%
16	350	68.92%	67.97%	67.26%
15	67	61.29%	63.77%	64.13%

Even though random forest is designed to boost the accuracy of Decision Tree by account for more variance, their performances are similar in these trials. Most of the splitting values are binary, therefore, Decision Tree with binary split on Data 15 returns the same result, but if they are transformed into binary, Random Forest returns higher with 64.56%.



Regarding precision and recall, MNB works better than the trees. Ashari et. al (2013) explains because it does not follow any loss function to penalize the misclassification. It simply rely on high probability assigned to the correct class and it accounts the distribution of attributes dependencies over classes.

A good classifier should have balance high precision and recall for all classes, and meaningful results. Therefore, to narrow down the research, Data 16 is chosen as the closest one since it is stopwords-free with more even precision and recall. One way could be done to improve this model is by adding other attributes. Atkinson (2018) classify some essential keywords in tweet text into four identities.

Racial/Ethnic	Sexual Orientation	Religious	Candidate Preferences
"Black", "White", "Native American", "Asian", "Mexican", "Hispanic", "Latino"	"LGBT", "Bisexual", "Homosexual", "Heterosexual", "Male", "Female"	"Muslim", "Christian", "Jew"	"4Trump", "forTrump", "4Hillary", "forHillary"

Only *"black"* ,*"white"*, and *"muslim"* exist in the data. By adding the others, the accuracy went up slightly from 68.91% to 68.97% with the majority error is Trolls identified as Other (19% of total).

Tweet Users	Precision	Recall	F-Score	ROC Area
Right Troll	0.824	0.499	0.622	0.804
Left Troll	0.668	0.560	0.609	0.798
Other	0.650	0.979	0.781	0.899
Average	0.714	0.690	0.675	0.836

Recall the behaviour of trolls according to retweet and the corresponding words to both trolls. By transform the class into simply Trolls and Other, the accuracy rise sharply to 77.63% with the majority error is Other predicted as Trolls (20% of total).

Tweet Users	Precision	Recall	F-Score	ROC Area
Trolls	0.759	0.956	0.846	0.898
Other	0.852	0.454	0.592	0.898
Average	0.792	0.776	0.755	0.898

If the prediction is made in a user basis using a collection of tweet text then the evaluation need to be extended. User can be classified based on its majority tweet predictions. Thus the accuracy in user-level using the three-classes model is 76% with the following confusion matrix.

Actual User Types	Tweet-Level Predictions			User-Level Predictions		
	Left Troll	Right Troll	Other	Left Troll	Right Troll	Other
LeftTroll	9727	1790	5862	5	1	1
RightTroll	4626	9337	4743	4	7	4
Other	215	199	19695			20
Total	14568	11326	30300	9	8	25

And the prediction result using this model for the test data is as follows.

Actual User Types	Predictions	
	Tweet-Level	User-level
LeftTroll	12622	9
RightTroll	10517	10
Other	20713	14
Total	43852	33

V. CONCLUSION

In conclusion, only to answer the research question, it is obvious that some essential terms in tweet text are helpful to predict trolls with a certain degree of misclassification depends on how rigid typical trolls need to be predicted and the purpose of the prediction itself. Whether a predictive model is implement to detect trolls from a new tweet text or a collection of tweet text has different evaluation.

REFERENCES

Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. International Journal of Advanced Computer Science and Applications (IJACSA), 4(11).

Atkinson, Ryan. (2018). Content Analysis of Russian Trolls' Tweets Circa the 2016 United States Presidential Election.Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/199858.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data engineering, 15(6), 1437-1447.

Linvill, Darren and Patrick Warren (2018) Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building (working paper). Clemson University.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48).

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). New York: IBM.

Tayade, P. M., Shaikh, S. S., & Deshmukh, S. N. (2017). To Discover Trolling Patterns in Social Media: Troll Filter.