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

1. Introduction

With increasing penetration of Internet in this so-called information era, human perception and understanding become inextricably entwined with information online. Troll industry emerges as a network industry chain where employees known as 'troller' sow discord by posting inflammatory and digressive, extraneous, or off-topic messages online via numerous fake social accounts [1]. It is now widely used to deface cerebrities by posting negative comments or even infiltrate national politics to maliciously provoke others. In June 2018, US released a list of human-operated troll accounts associated with the activity of Russia's Internet Research Agency (IRA) and all tweets published by these accounts from 2015 to 2017 were collected and analysed. This report aims to apply Machine Learning Methods including KNN and Naive Bayes Classifiers over the collected dataset to help identify trolls on Twitter and explore how trollers pretend themselves as a normal Tweet user.

In this report, classifiers are designed and implemented to distinguish whether a tweet is more likely to be a 'left troll', a 'right troll' or 'others'.

- Left troll: socially liberal message about gender, sexual, religious, and racial identity with the intent to attack mainstream Democratic politicians. It can be concluded as an opposition with a radical tone.
- Right troll: nativist and right-leaning populist messages. It can be concluded as supported remarks mimicking stereotypical Trump supporters.
- Others: three categories including News Feed (overwhelmingly presenting themselves as news aggregators by linking to legitimate news and issues of local interest), Hashtag Gamer and Fearmonger (a fake news probable as a trial run for Russians to gauge how easily they could influence Americans online) [2].

2. Dataset

The dataset, as curated by [3], has been adapted from a much larger dataset of tweets from users associated with the Internet Research Agency, a Russian company engaging in troll farm writing employing fake accounts registered on major social networks or other media. The dataset comprises 223k tweets from 175 users and was partitioned into three parts: training data used to build models, dev data to evaluate models and test data, a set of unlabelled data, to finally test the models. All datasets have been pre-processed, denoted by 'Best type', by recording the term frequency and selecting a subset of the terms with the greatest Mutual Information and Chi-Square values. Each includes 56194 instances (17379

actual left trolls, 18706 actual right trolls and 20109 actual other). All the datasets were preprocessed by removing attribute 1st 'tweet-id' and attribute 2nd 'user-id'.

3. Evaluation metrics

Throughout the report, the following metrics will be applied to identify how well a model identifies the trolls on Twitter:

- Confusion matrix: a table used to describe the effectiveness of a classification model on a set of test data for which the true values are known [4]. The following table shows confusion matrix for multiple classes:

		Predicted Number			
Actural Number		Class 1	Class2	...	Class n
	Class1	x_{11}	x_{12}		x_{1n}
	Class2	x_{21}	x_{22}	...	x_{2n}
	...	...	...	...	
	Class	x_{n1}	x_{n2}		x_{nn}

Table 1: Confusion matrix for multiple classes

- Precision (Pi): proportion of positive(i) predictions that are correct, also called positive predicted value. As the ultimate goal is to find trolls from all tweets, the system can be assumed to have a certain tolerance for false positive predictions

- Recall (Ri): accuracy with respect to positive(i) cases also called true positive rate. The recall of left troll will be a key observation to judge political assaults how well the system picks up political assaults from countless tweets on Twitter.

- Accuracy (ACC): the proportion of instances for which we have correctly predicted the label. Accuracy is a good measure when the target variable classes in the data are nearly balanced. It should not be as a measure alone when the target variable classes are a majority of one class [4].

4. Methodology

4.1. K-nearest Neighbour classification

k nearest neighbor classification or (kNN) assigns the majority class of the k nearest neighbors to a test document [5]. For a KNN model, the aspects we normally need to decide including:

- Parameter k: k can be selected from 1-n, often based on experience or knowledge of the problem to be solved. k is desirable to odd to make ties less likely except k = 1 (not very robust). k = 3 and k = 5 were experimented separately in this system.
- Nearest Neighbour Search Algorithm: decide distance function and data structure which attempt to reduce distance calculations. LinearNNsearch and BallTree both with Euclidean distance were implemented.

4.2. Naïve Bayes (NB) Classifiers

Naïve Bayes classification [2] is based on Bayes' probability rule and Naive Bayes Conditional Independence Assumption, say assuming the probability of observing the conjunction of attributes equal to the product of the individual probabilities P (xi|cj). Thus, we can classify instances by selecting one with maximum posterior probability. Two NB classifiers, NB classifier for Binary class and NB multinomial classifier were implemented. With a multinomial model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial [6].

All the models were experimented on Weka 3, a data mining software in java.

5. Results

Below are the precision and recall of each class (L for left troll, R for right troll and O for other) resulted from 6 classifiers and confusion matrix for NB multinominal classifier and KNN for k = 5. All classifier accuracy are very similar ranging from 61.5% - 64.7%, which indicates decent effectiveness of these classifiers. Variance of k value and search algorithm have little effect on results. NB Multinominal classifier returns a highest recall of Left troll (0.56) but very low recall of Right troll, which present a good effectiveness on picking up left trolls on Tweeter but bad on right troll.

	P(L)	P(R)	P(O)	R(L)	R(R)	R(O)	ACC
Knn k=3 linear	0.613	0.669	0.628	0.445	0.447	0.971	63.4036 %
Knn k=3 ball tree	0.613	0.669	0.628	0.445	0.447	0.971	63.4036 %
Knn k=5 linear	0.602	0.668	0.639	0.466	0.436	0.970	63.6313 %
Knn k=5 ball tree	0.613	0.669	0.630	0.445	0.450	0.971	63.4747 %
NB	0.559	0.644	0.635	0.499	0.343	0.970	61.5511 %
NB(multinominal)	0.610	0.782	0.626	0.560	0.380	0.970	64.6617 %

Table 2: Results of 6 classifiers for 56194 instances (17379 actual left trolls, 18706 right trolls and 20109 other)

		Predicted Number		
Actual		Left troll	Right troll	Other
	Left troll	9724	1598	6057
	Right troll	5994	7106	5606

	Other	217	386	19506

Table3: Confusion Matrix result for NB Multinominal classifier

		Predicted Number		
Actural Number		Left troll	Right troll	Other
	Left troll	7735	3767	5877
	Right troll	4682	8410	5614
	Other	191	394	19524

Table 4: Confusion Matrix result for KNN (k=5, BallTree)

6. Evaluation

Super high recall of other trolls (the value of R(O) from all the class is no lower than 0.97), which means nearly all 'other' Tweets can be found out and very few tweets which were predicted as 'left' or 'right' trolls are actually 'other' tweets. It shows that 'other' tweets greatly depend on term frequency (no matter calculated from which classifiers), that is, most 'other' tweets contain groups of fixed words or even sentences. This is probably because trollers tend to copy 'other' tweets from an 'other' text collection, as listed before, related to News Feed, Hashtag Gamer and Fearmonger. Unlike left or right trolls created by a large cost of human capital to a troll factory, trollers through the 'other' text collection can much less costly turn Twitter homepages into one with American news (several are fake) and harmless online activities as if the Twitter accounts belonged to real American citizens. Behaving like MAGA Americans, their Tweeters would be interspersed with political leanings seamless and cleverly. Note that Precision(other) is 0.63 of tiny float for all applied classifiers, and the Confusion Matrix shows that about 1/3 of pretended 'other' tweets are actually left trolls or right trolls. This may because some of 'Other' tweets that trollers write follow fixed patterns implied politics or because some left or right trolls not just objective information but presented in tone which cannot be distinguished literally.

7. Conclusion

We can effectively use tweet text to help us to identify majority of trolls on Twitter based on term frequency. Most 'Left trolls' and 'Right trolls' can be picked up and those not classified correctly potentially due to misclassification to 'other' when both left or right trolls and 'other' have similar term combinations, or to undetectability of trolls with sentiment, tone or implied meanings. 'Other' tweets including news link, tag game record and fake crisis, however, can be nearly perfectly classified, probably because they are generated from a same text collection. Through it, trollers may pretend a Tweet account as the one belonged to a citizen with one of all identities.

References

[1] A. Haque. Unsupervised Learning and Online Detection of Internet Trolls in Streaming Data.

[2] O. Roeder, "We Gave You 3 Million Russian Troll Tweets. Here's What You've Found So Far.", *FiveThirtyEight*, 2018. [Online]. Available: https://fivethirtyeight.com/features/what-you-found-in-3-million-russian-troll-tweets/. [Accessed: 10- Oct- 2018].

[3] Linvill, Darren and Patrick Warren (2018) Troll factories: *The Internet Research Agency and state-sponsored agenda building (working paper)*. Clemson University.

[4] Performance Metrics for Classification problems in Machine Learning", *Medium*, 2018. [Online]. Available: https://medium.com/greyatom/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b. [Accessed: 10- Oct- 2018].

[5] "A Complete Guide to K-Nearest-Neighbors with Applications in Python and R", *Kevinzakka.github.io*, 2018. [Online]. Available: https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/. [Accessed: 10- Oct- 2018].

[6] Brownlee, "Naive Bayes for Machine Learning", *Machine Learning Mastery*, 2018. [Online]. Available: https://machinelearningmastery.com/naive-bayes-for-machine-learning/. [Accessed: 10- Oct- 2018].