School of Computing and Information Systems
The University of Melbourne
**COMP90049 Knowledge Technologies, Semester 2 2018**
Project 2: Twitter Trolls and the Tweeters who Love them

## 1 Introduction

Machine learning now is one of the most prevalent concepts in computer science. If machine learning algorithms are able to automatically recognizes trolls on Twitter, it will significantly increase system efficiency and reduce labor costs. Therefore, the objective of this report is to demonstrate the feasibility of that applying several machine learning algorithms to identify troll users based on content of their tweets. In order to contextualize and evaluate relevant algorithm, a brief description of datasets and evaluation metrics will be introduced. Subsequently, the report presents the outcomes of naïve Bayes and K-NN in terms of evaluation metrics and illustrative examples. Finally, the argument of whether machine learning algorithms can effectively identify twitter trolls will be discussed.

## 2 Dataset

The dataset has been adapted from a much larger dataset of tweets (Linvill, Darren & Warren, 2018). There are several data files which have distinct size and format. The whole dataset comprises about 3M tweets, issued by 2848 unique users. For increasing running speed, a smaller dataset which includes 223K tweets is constructed and partitioned into training (60%), development (20%), and test (20%) sets. In addition, it is notable that all characters that are not English alphabetic are removed in order to avoid consequently garbled content.

## 3 Evaluation Metric

### 3.1 Overall Accuracy

Accuracy is able to illustrate the overall performance of models clearly by counting the proportion of correctly classified instances (Blum & Langley, 1997).

### 3.2 Precision & Recall

Precision is the proportion of true positive instances in predicted positive instances, while Recall is the proportion of true positive samples in all positive samples (Blum & Langley, 1997). A high recall rate means the classification performance of this category is better.

### 3.4 F-Measure

Precision and recall are a pair of contradictory metrics (Blum & Langley, 1997). F-measure combines the recall rate with the precision rate, allowing to express different preferences for the precision/recall rate by coefficient $\alpha$ (When $\alpha > 1$, the recall rate has a greater impact, and when $\alpha < 1$, the precision has a greater impact).

## 4 Literature Review

### 4.1 Naïve Bayes

Naive Bayes algorithm is a type of probabilistic classification method based on Bayes' theorem. In order to simplify the calculation, this algorithm assumes the distribution of each features are independent (McCallum & Nigam, 1998). Additionally, it merely requires a small dataset to estimate necessary parameters.

Although independent assumptions are generally inaccurate in fact, naive Bayes classifiers still achieve remarkable results in practice (Rish, 2001). As can be seen, decoupling between various conditional features means that the distribution of each feature can be independently estimated as a

one-dimensional distribution, which effectively avoids the dimensionality hazard and allows classifiers to be more concise and clearer.

### 4.2 K-Nearest Neighbours

K-NN is an instance learning method, where all classification missions are deferred until test stage. In addition, k-NN also support to modify the window-size, which means that unlabelled instances are able to be approximated locally (Wu et al., 2007). In the classification phase, user ought to select a distance calculation method, and then an instance (Vector) can be classified a label which is most frequent among the top-k (k is a user-defined constant) nearest instances.

K-NN algorithm is among the simplest method of all machine learning algorithms, which is easy to understand and implement (Wu et al., 2007). However, the accuracy of K-NN would be seriously affected by major features. Moreover, and the time complexity and storage space of k-NN algorithm increase rapidly as the training dataset size and feature dimension increase.

## 5. Methodology

### 5.1 Feature Selection

Conceptually, Feature selection is of importance in classification as features have effect on final outcome to a great extent. There are numerous features, which have prominent frequency but likely to be pointless for classification (such as "an", "of" and "you", etc.) (Chen, Huang, Tian & Qu, 2009). Therefore, removing these features by TF-IDF value may enable classifiers acquiring higher performance (Aizawa, 2003).

Additionally, "Tweet-id" and "User-id" may not be helpful for classification model since every instance has unique id values in the train dataset and development dataset.

### 5.2 Naïve Bayes

According to the conditional independence formula, if X and Y are independent of each other, then:

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

Subsequently, Bayes formula can be derived by the above formula:

$$P(Y_k|X) = \frac{P(X|Y_k) * P(Y_k)}{\sum_k P(X|Y = Y_k) * P(Y_k)}$$

Obviously, $P(X|Y = Y^k)$ is a complicated conditional distribution with k dimensions. For this, naïve Bayes model makes a bold assumption that the n dimensions are independent of each other. So that the naïve Bayes formula can be derived:

$$C_{result} = argmax(P(Y = C_k) * \prod_{j=1}^{n} P(X_j = X_j^{test}|Y = C_k))$$

#### 5.2.1 Analysis

All training of naïve Bayes was done on the Best-200 dataset for more accuracy and avoiding under-fitting. The following table 5.1 has shown that the performance of naive Bayes in distinct situations. As can be seen, naïve Bayes model does not have a promising performance on the original dataset (Merely 50.1%).

From my own perspective, "User-id" is a fairly decent feature in the same data set (Since most instances which have the same user-id seem also having the same class) but dev dataset and training dataset do not have duplicated "User-id", which means these features are pointless and even has negative effect on the outcome. Consequently, the performance of naïve Bayes has been significantly improved after removing "Tweet-id" and "User-id", in terms of overall

| Feature situation | Overall Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Original Dataset | 50.1086 % | LT: 33.1 %<br>RT: 46.2 %<br>Other: 98.0 % | LT: 55.9 %<br>RT: 37.5 %<br>Other: 56.9 % | LT: 41.6 %<br>RT: 41.4 %<br>Other: 72.0 % |
| Tweet-id and User-id removed | 73.0772 % | LT: 62.8 %<br>RT: 80.0 %<br>Other: 77.8 % | LT: 70.1 %<br>RT: 50.4 %<br>Other: 96.8 % | LT: 66.2 %<br>RT: 61.8 %<br>Other: 86.2 % |
| Pointless Words removed | 73.2053 % | LT: 63.6 %<br>RT: 80.1 %<br>Other: 77.2 % | LT: 69.5 %<br>RT: 50.1 %<br>Other: 97.0 % | LT: 66.4 %<br>RT: 62.3 %<br>Other: 86.0 % |

*Table 5.1: The performance of naïve Bayes under the Best-200 Dataset*

accuracy (From 50.1086% to 73.0772 %) and F-measure (Up 14.6% for left trolls, 20.4% for right trolls and 14.2% for others respectively).

Moreover, it is notable that removing low TF-IDF features does not have obvious benefits in this case. For this, a possible reason is that all the features in best-200 dataset are high-weight features which have been filtered, so filtering features by TF-IDF have not achieved the expected effect (Aizawa, 2003).

### 5.3 K-Nearest Neighbours

As a non-parametric statistical method, the implement of k-NN is simply constructed (Wu et al., 2007). Given distance measures, the k points nearest to x can be found in the training dataset T, and the neighborhood of x which containing k points is denoted as N(k). Later, the class Y of X in N(k) is determined by the classification decision rule (such as majority vote), and the formula is as follows:

$$y = argmax_{c_j} \sum_{x_j \in N_k} F(Y_j = C_j)$$

Where F is an indicator function, F equals 1 when the equation in parentheses is true, otherwise equals 0.

#### 5.3.1 Analysis

Similarly, all training was done on the Best-200 dataset. As parameter k based on experience, the final value of k is generally determined by manual test (Wu et al., 2007). Where K equals 5, k-NN model obtains an acceptable result on the dataset (Overall accuracy is 60.9513%).

Comparing with Recall rate of "Left trolls" and "Right Trolls", the value of "Other" is obviously higher, which might be caused by the amount of "Other" class dominates the whole dataset, indirectly resulting in the result of k-NN prone to this class.

## 6 Conclusion

This report presents and analyses naive Bayes and k-NN algorithms. Through experiments in WEKA, both naïve Bayes model and k-NN algorithm have a decent performance on the tweets dataset (Best-200 in this case). Furthermore, Naive Bayes model obtains an outstanding outcome that overall accuracy is over 73%. According to the outcome of above algorithms, the Tweets data has been shown to be useful for identifying Trolls. In addition, we have noticed that the "Other" class had higher F-measures in the results, a possible reason is that the amount of "Other" is more than other two classes in the dataset.

## Reference


Aizawa, A. (2003). An information-theoretic perspective of TF-IDF measures. *Information Processing & Management*, *39*(1), 45-65.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, *97*(1-2), 245-271.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3), 5432-5435.

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 naïve 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.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.