Coursework 1: Predictive Data Mining Group Project

Project Brief

Group Name: *Troublemaker*

1. Team Members

We are team of five members as following:

	Full name	Email
1	Lu Yang	ly4n17@soton.ac.uk
2	Jinmu Liu	jl12m17@soton.ac.uk
3	Jiachen Li	jl4g17@soton.ac.uk
4	Qin Yu	qy2a17@soton.ac.uk
5	Yue Wang	yw3y17@soton.ac.uk
6	Junjie Lu	jl9n17@soton.ac.uk

2. Description

Description: The toxic comment classification problem is from Kaggle which provides various data mining competitions. We aim to detect and classify the offensive comments by their contents. Firstly, we are planning to do the binary classification and distinguish whether toxic or not. After we get the result, we will continue to explore the extent of the toxic comments and determine its category (see Data Section). The main algorithms we plan to use are Naive Bayesian classification, random forest, neural network and support vector machine. Finally, we will compare the results of these algorithms and give an optimal method to implement multi-classification of toxic comments.

3. Data

Training set: The training set contains 159571 observations and 8 columns (id, comment text, toxic, severe toxic, obscene, threat, insult, identity hate) which marks the comment in the last six columns with 0 or 1 (figure 1), which means the comment is considered to be offensive or not or which kinds of offence. And the comments can belong to multiple categories of offence.

	A	B	C	D	E	F	G	H
1	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
56	001ffdcc3e7fb49c	Awesome! Then I'll sim	0	0	0	0	0	0
57	0020e7119b96eeeb	Stupid peace of shit st	1	1	1	0	1	0
58	0020fd96ed3b8c8b	=Tony Sidaway is obvic	1	0	1	0	1	0
59	00218d74784ce50b	"	0	0	0	0	0	0

Figure 1: Training set

Test set: The test set is consist of 153164 observations and 2 columns, the id and text.

3. Git Repository

https://github.com/yeunglo/DataMining-Group-Project

4. Kaggle Challenge

https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge