1 *Introduction*

The machine learning algorithm we ran for this assignment are supervised learning. This is because the correct answers in terms of classified and unclassified tweets have already been given. In terms of which classification algorithm to use, heuristically speaking this is a classification problem so linear regression may not be the best because these types of regression would imply to work best for continued output values. Therefore, our hunch is to use logistic regression with another choice between SVM and Naives Bayes.

The approach is to take our classified tweets and feed them to one of the learning algorithms and from this we will get a hypothesis. The job of the hypothesis is then to take input features [1] and then try to predict whether the tweets are positive or negative.

[1] in our case tweet words that may be useful for classification

2 *Choice of Learning Algorithm*

For this assignment, logistic regression and support vector machines were used. The convergence of Naive Bayes algorithm was very slowly and given the large data set, this algorithm was avoided.

For logistic regression, variations in regulization (to avoid overfitting) and type of performance metric (precision, accuray etc) did not change the performance of the algorithm. Also use of GridSearchCV from the sklearn module did not result in any meaningful change compared to straight implementation of the algorithm. Changing the C parameter for regularization again did not make any meaningful difference in the output. Fig. 1 adaptations of the logistic algorithm with changing samples.

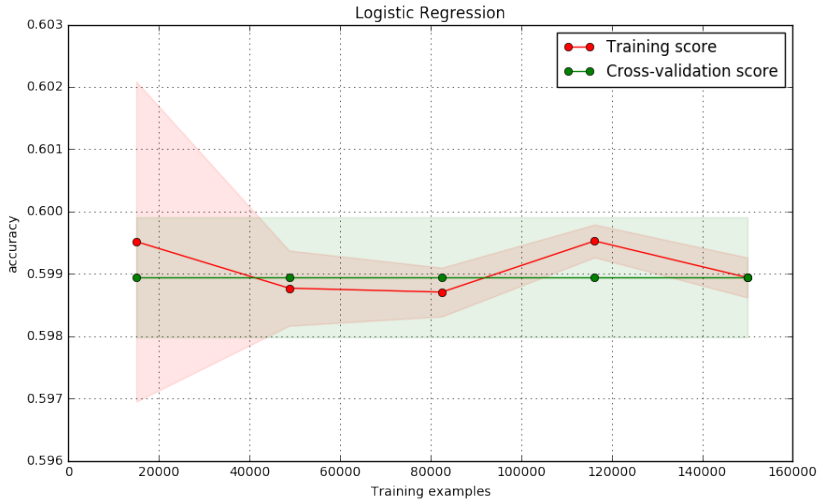

Figure 1: Logistic regression learning and cross validation.

SVM algorithm was much more challenging to implement. First of all, for any large test set, the algorithm did not converge at all. As a result, we need to reduce the training size, use a linear solver, lower the value of C or a combination of these. Although GridSearchCV was used, it is not very appropriate for this type of basic linear combination. If we had multiple

3 *Functions used*

The same functions as those of Assignment 1 were reused. Each tweet was given a score based on the value of the words it contained as given by the corpus file.

4 *Performance of Algorithms*

Table 1 shows the perofrmance metrics for the two algorithms implemented. The data is almost identical in both cases.

Metric	**Logistic**	**SVM**
Accuracy	0.599633	0.5993
Precision	0.65857	0.912
Recall	0.40313	0.40496
F1 Score	0.50012	0.50253

Table 1: Performance Metrics for Two learning algorithms

In terms of commenting on the actual algorithm, the accuracy is at 60% which is slightly better than guessing so not abysmal but not very accurate either. The precision score is higher at 67% and 91%. This is encouraging, it essentially means for 9 out of 10 tweets, if the SVM classifies our result as positive, then it is positive. This combined with a low sensitivity value indicates that using positive tweets is a good measure for determining insight into the elections.

5 *Insight into the elections*

Our algorithms did not perform extremely well, but the positive tweet seem to be better classified and also no overly optimistic. This means that if our data can show a correlation between a party and positive tweets, then we can say that the party is likely doing better than others.

Using this, the logistic regression algorithm was run on the unclassified tweets[2] . If the tweets were positive, they were then checked to see if they match one of the parties.

As an be seen, the Liberal party seemed to have the most number of "positive" tweets as determined by the algorithm [3] followed by the

[2] The SVM algorithm took much longer to converge and was not used.

[3] which is only correct 60% of the time but does better for positive tweets.

Party	Number of Positive Tweets	Percentage of seats in Parliament
Liberal	35	56.2 %
Conservative	9	30.3 %
NDP	2	13.5 %

Table 2: Positive tweets associated with each party (for unclassified tweet set)

Conservatives and NDP. As we know, the election results also were a Liberal majority, so this primitive analysis would have worked in predicting the results of the election.

6 Conclusion

Some concluding remarks:

1. This assignment borders artificial intelligence and the method for scoring tweets as positive or negative was rather primitive.
2. As a result, prediction with this algorithm are rather suspect. Yes, using the positive tweet sentiment we achieved a result that matched the election but repeatability is suspect.
3. Given the one feature size, logistic regression was the most appropriate for this analysis. To use other algorithms such as Naives Bayes would require multi-feature data set which was too time consuming to develop for this work.