BigData&DataAnalytics

Learnings lessons of Development of sentiment analysis

Mauricio Carvajal
IOT ANALYTICS

Table of contents

1. Introduction

In this section we have the summarize all the learnings obtain thru the development of the project. The lessons learnings start from the preprocess fo the data, with start with the features selection, starting with the finding the correlation, examinating the feacture variance, the recursive feacture elimination, also with engineering the dependen variable and principal component analsys, all of this has been tested to the different models (Randon Forest, Support Vector Machine, KNN and C5.0 in order to choose the best model, and then run do the predictions that are require for this sentimental analsys.

In summary this documents is divide in 3 principal sections:

1. Learnings on the choosing the clasiffier and feacture selection for trainings the models
2. Summary of all the activites involve
3. Learnings for future projects

2.1 Classifier and features selected

The First step is to close the classifier, so for this task 4 algoriths has been trained and evaluated as shown in the table 1.

In the talbe we have the results for Iphone and Galaxy, also the classifier that has been used, and how we evaluated the model. The first step is about test Time, due we already is setting pararelization in our machine, this play important rol. Also is the evaluation of the performance of the different models, betten how is the evaluation of the model itself, and how is performance of the testing data sets.

- C5.0
- Random Forest
- SVM (from the e1071 package)
- kknn (from the kknn package)

Device	Classifier	TestTime	Model		Testing	
			Accuracy	Kappa	Accuracy	Kappa
Iphone	C5.0	1.09	0.76099	0.5366	0.7528	0.5214
	RF	7.03	0.7618	0.5433	0.7580	0.5358
	SVM	14.14	0.7216	0.4443	0.7216	0.4443
	Knn	4.19	0.7674	0.5504	0.7674	0.5504
Galaxy	C5.0	5.70	0.7128	0.4758	0.77023	0.5416
	RF	77.99	0.76409	0.5302	0.7742	0.5534
	SVM	55.78	0.7130	0.3923	0.7130	0.3923
	Knn	13.09	0.7642	0.5298	0.7642	0.5298

It’s important to review that the model that has been choses is **c5.0**, but this is not the one that has be best performance in terms of Accuracy and Kappa, but it much faster thatn the other algos, and the performance is less than 5%. For further details please look the full technical document and the script that is already attached.

2.2 Features selection

Once we are clear in the Classifier that we are going to use, then it's needed to show how the data is going to be preprocess, in this cas e we have 4 different methos that are shown in the table 2.

Device	Datasets	TestTime	Accuracy	Kappa
Iphone	Recursive Feature Elimination	1.27	0.7237	0.5018
	Correlation Elimination	2.56	0.7467	0.5323
	Feature Variance(NZV)	1.36	0.7580	0.5246
	RC	7.22	0.7997	0.5630
Galaxy	Recursive Feature Elimination	5.70	0.7128	0.4758
	Correlation Elimination	3.54	0.7135	0.4738
	Feature Variance(NZV)	1.17	0.7482	0.4883
	RC	3.47	0.7808	0.5386

2.3 Comparative performance

In the following table the results are been shown, do be easy compare. It's important to mention that both are using the same model and features due to one important parameter that has been set is the testime.

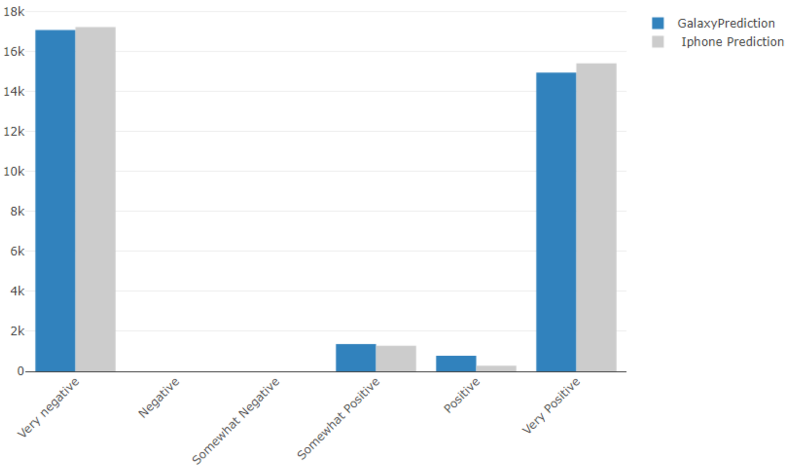
Between Iphnoe an Galaxy, the results regarding performance is very similar, so a least this give us the advantage compare those, but as it's explained for future works is needed to work on the datasets to have better results.

For further details please look the full technical document and the script that is already attached.

Device	Final model	TestTime	**Model**		**Training**	
			Accuracy	**Kappa**	**Accuracy**	**Kappa**
Iphone	C5.0 Feature Variance(NZV)	1.36	0.7580	0.5246	0.7542	0.5159
Galaxy	C5.0 Feature Variance(NZV)	1.17	0.7482	0.4883	0.7542	0.5159

2.5 Prediction

The following picture shows the results side by side of the predicitons of the galaxy and the iphone. In whis is show that both present similar behavior, but galaxy has a trend to have more positive than iphone that present a love/hate relationship.



2. Summary

2.1 What worked well.

Following the metolodgy seems to be a good approach to develop a model to predict sentiment analsys. The data has been preprocessed with different methods, then use “Out of the Box Method” help us to ensure that we chose the best classifier. In this particular project, due to this is late. The Test Time was very important, so we optimized using parallel cores of the computers, and evaluate with algo was more effective.

2.2 What didn’t work.

The accuracy of the predioons can be improved, even for this project that the data has been preproced of different ways and used multiple classifiers, the results are below of the 78%.

2.3 What was difficult

The difficult part was to try to have performance indicates of the predictions that has been done, in this case did it for iphone and galaxy, but depends of the model and the preprocess of the information, it’s hard to evaluate which perform better.

3. Next projects

For future projects it’s suggested to maintin the “Out of the box model”, due to it has very good results, and allows to have bigger overview on the best clasiffiers. But as area of improvement the large data sets needs to be more pre process, do have more relevant information, I suggest to analysis from the beginning the distrituion, and the relationship between the datasets that the model is devolp and tested and the real data set, due to in this case those are quite difference. So this can be improved with more pre process, or more time on the development of both data sets.