COMP90049 Project 2 Report: Geolocation of Tweets with Machine Learning

Name : Akash Singh

1. Introduction

With the popularity of online social media in recent years, the amount of user data available for data mining has led to the development of efficient approaches to determine users geographical location. Twitter a microblogging service has a very wide variety of user demographics base in terms of geographical location. The users tweet contains a variety of words, some of which are related to a particular geographical region.For example, '*howdy'* is an informal greeting way in Texas, US. With geographically diverse user base, Twitter generates a very huge amount of data which can be used for developing machine learning techniques and models that can detect a user's location by analysing only the tweet words with high precision and accuracy. For example, the paper Web-a-Where: Geotagging Web Content [4], achieves 82% of precision.

In this report, three major machine learning algorithm has been examined i.e, ZeroR as baseline algorithm, NaiveBayesMultinomial and Decision tree and one ensemble machine learning algorithm Random Forest to develop classifiers that can predict a users location from analysing only tweet contents. The intuition (hypothesis) behind this approach is that user tweet content may contain location specific content - either specific place name or phrase likely to be specific to a particular location[1].

2. Dataset

The dataset examined and utilized in this report is presented in Eisenstein, Jacob, et al.[2] and Rahimi, Afshin, Trevor, and Timothy [3]. A sample representation of the raw data in form of ARFF file format (Table 1) has been used to develop the classifier.

File	Number of Instance
Training	96585
Development	34028
Test	32977

Table 1

3. Literature Review

The paper Web-a-Where: Geotagging Web Content [4] discuss a methodology for geo-tagging a web-page based on mentions of the places in the page content, which appears quite effective way of localizing the web-page to a place or region achieving 82% of precision.

Geolocating blogs from their textual content [5], utilizes user's post content purely to determine their location. It uses the place name mentions in a blog to determine an author's location. Achieving an accuracy of 63% on a collection of 844 blogs with known locations.

These research work clearly show that microblogging applications like Twitter have enough real-time human sensor data, which can be utilized effectively with adequate accuracy and precision to determine author's location.

4. Methodology

4.1 Attribute Filtering

In order to develop a classifier with better precision, a heuristic approach has been taken to filter few attributes on the following reasons:

1. Very common and general english language words and also they classify all the labels with almost equal probability. For example, words like and, the, are, dead, that, etc. Image 1 shows the probability using

NaiveBayesMultinomial with cross-validation 10 fold on train-best10.arff file.

		The probability of the given class	
	NewYork	California	Georgia
and	0.13	0.17	0.17
are	0.03	0.06	0.05
the	0.3	0.36	0.36

Image 1

2. Internet slang words like lamoo, lml, lol etc has been filtered out because they are very popular and use by diversified users, based on any location on the globe and hence they contribute nothing more than noise to the training data and increases the complexity and variance of the developed model on this data. This is the author's own intuition and knowledge.

3. Tweet words like atlanta, la, gw, famu,etc appeared very important which in turn actually associated with some event happened at a particular location or they are associated with some entity or they are words of other languages which signifies the concentration of people associated with the language to a particular geo-region. Image 2 shows the probability using NaiveBayesMultinomial with cross-validation 10 fold on train-best10.arff

		The probability of the given class	
	NewYork	California	Georgia
atlanta	0	0	0.01
la	0	0.01	0
gw	0	0.01	0

Image 2

and these are the main attributes which help to better train the classifies with high precision.

The final list of refined attributes used here is: "*atl, atlanta, childplease, famu, gsu, gw, hella, inhighschool, la, parody, thatisall, wet*".

4.2 Test-Option selection

Two test-options has been examined using NaiveBayesMultinomial. First is percentage-split 90-10% and other one is cross-validation with 10 folds, Image 3&4 confusion matrices show that classier with cross-validation test-option is better able to classify the instances to their respective correct labels, hence developing better learned classifier.

Percentage-split 90-10 %

	confusion matrix		
a	b	c	classified as
54148	481	46	a=NewYork
17448	1041	23	b=California
13314	188	237	c = Georgia

Image 3

Cross-validation 10 fold

	confusion matrix		
a	b	c	classified as
60551	194	65	a=NewYork
20100	489	40	b=California
14708	63	375	c = Georgia

Image 4

So throughout the development of the different classifiers cross-validation has been used.

4.3 Evaluation metric

The primary aim of this report is to develop a classifier which can give good prediction of tweeter's location. Accuracy and Precision are the main focused evaluation metric.Reason for selecting Precision is because, first it tells how many times actually classifier is true out of total true prediction for a single label. Second Precision's division factor is only the number of times classifier predicts true for a particular label and as the number of instances are very high so metric like Recall is not a very appropriate choice. Though Recall and F-measures are taken into consideration.

4.4 Classifier Algorithm

Different machine learning algorithms is taken into consideration and they are ZeroR as baseline, NaiveBayesMultinomial, Decision tree, Random Tree. Further analysis on each of them is presented in the next section.

5. Critical Analysis

Cross-validation 10 fold on train-best10.arff & dev-best10.arff files is used to develop results with different ML algorithms.

5.1 ZeroR as baseline

This ML algorithm works on majority rule hence classifier will be trained to only assign label that has the highest number of instances. Image 5 shows only NewYork has been assigned as label predictions to instances using ZeroR.

Precision		Recall		F-measure		Class
Trainig	Dev	Training	Dev	Traing	Dev	
0.63	0.644	1	1	0.773	0.784	NewYork
?	?	?	?	?	?	California
?	?	?	?	?	?	Georgia

Image 5

The accuracy achieved for training and development data is shown in Table 2.

Accuracy	Data
62.9601	Traning
64.4146	Dev

Table 2

5.2 NaiveBayesMultinomial

This ML algorithm works on probability, in other words it tries to estimate probabilities for each class and picks a class with the maximum probability i.e most likely class is taken into consideration by this algorithm. For example in raw training tweets file there are 894 occurrences of attribute '*inhighschool*' given class NewYork, 42 given class California and 56 given class Georgia and this is clearly seen in Image 6, the algorithm gives highest probability to NewYork than California and Georgia:

	The probability of a word given the class		
	NewYork	California	Georgia
inhighschool	0.74	0.09	0.13

Image 6

Evaluation metric Image 7, shows that classifier now can give predictions to other classes also i.e California and Georgia as compared with the prediction results of ZeroR classifier.

Precision		Recall		F-measure		Class
Trainig	Dev	Training	Dev	Traing	Dev	
0.632	0.646	0.999	0.999	0.774	0.785	NewYork
0.762	0.632	0.009	0.007	0.018	0.013	California
0.793	0.747	0.017	0.013	0.034	0.026	Georgia

Image 7

The accuracy achieved for training and development data is shown in Table 3. Results are better than ZeroR classifier.

Accuracy	Data
63.339	Traning
64.6703	Dev

Table 3

5.3 Decision Tree

This ML algorithm looks at each one the attributes to make the best possible inference about label, considering all other factors which can influence the prediction of the label. Hence the classifier is more generalised and gives better predictions.

Evaluation metric in Image 8 shows better generalised predictions for all the classes than NaiveBayesMultinomial.

Precision		Recall		F-measure		Class
Trainig	Dev	Training	Dev	Traing	Dev	
0.635	0.649	0.998	0.999	0.774	0.785	NewYork
0.769	0.647	0.01	0.006	0.02	0.012	California
0.801	0.742	0.014	0.012	0.027	0.025	Georgia

Image 8

The accuracy achieved for training and development data is shown in Table 4. Results are better than NaiveBayesMultinomial classifier.

Accuracy	Data
63.512	Traning
64.91	Dev

Table 4

5.4 Random Forest

This ML algorithm works on collective prediction made by a group of individual Decision Trees and final outcome is the label which has most votes from these individual Decision Trees. This classifier algorithm outperforms the individual Decision Tree performance and gives better learned and generalised classifier.

Precision		Recall		F-measure		Class
Trainig	Dev	Training	Dev	Traing	Dev	
0.637	0.651	0.997	0.997	0.774	0.784	NewYork
0.781	0.649	0.012	0.012	0.023	0.024	California
0.812	0.747	0.018	0.013	0.035	0.025	Georgia

Image 9

The accuracy achieved for training and development data is shown in Table 5. Results are better than individual Decision Trees classifier.

Accuracy	Data
63.73	Traning
65.14	Dev

Table 5

Conclusion

Different ML algorithms has been examined in this report, NaiveBayes algorithm scales very well and gives a better performance than ZeroR but it only considers probability as basic method of prediction.

A better performance is seen in Decision Tree as it works on Mutual Information concept and also considers other factors influencing the prediction.

Random Forest an ensemble ML algorithm gives a better generalised model as it makes use of group of Decision Trees with final output prediction through voting mechanism.

This report covers limited ML algorithms, there are a plethora of ML algorithms which can give a better learned and generalised classifier, for example Support Vector Machines and Bagging.

And a lot of knowledge can be drawn from user data available on social media in today's time which can help many industries to grow their business and technology.

Reference

1. [1] Cheng Z., Caverlee J., and Lee K. You are where you tweet: A content-based approach to geo-locating twitter users. In CIKM'10, Toronto, Ontario, Canada, 2010. ACM.
2. [2] Eisenstein, Jacob, et al. A latent variable model for geographic lexical variation. Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2010.
3. [3] Rahimi, Afshin, Trevor Cohn, and Timothy Baldwin. Semi-supervised user geolocation via graph convolutional networks. arXiv preprint arXiv:1804.08049 (2018).
4. [4] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In SIGIR, 2004.
5. [5] C. Fink, C. Piatko, J. Mayfield, T. Finin, and J. Martineau. Geolocating blogs from their textual content. In AAAI 2009 Spring Symposia on Social Semantic Web: Where Web 2.0 Meets Web 3.0, 2009.