**COMP90049 Report**

# Anonymous

COMP90049, Introduction to Machine Learning, Semester 1 2021

1. **Introduction**

Social media data importantly

twitter data is frequently used to identify the unique characteristics and identities of geographical regions, including topics of

interest ,linguistic styles and dialects political views. Machine learning has been used widely to perform many predictions such as sentimental analysis, geo-location prediction and this report focuses on predicting geo-location of the user from USERID and raw tweets. The classifiers discussed here are zero baseline classifier , KNN classifier, Naive Bayes classifier and different comparisons among different classifiers are being discussed here .This study is performed on twitter raw data and it consists of user id and raw tweets. Data is divided using holdout strategy: the data collection has been split into three parts: a training set, a development set, and a test set. Some feature engineering is done to the raw tweets and provided representations are discussed here :

First feature is where we have full raw tweets. Second type to feature is where we have we filtered the words in the data set, removing very frequent and very infrequent words and mapped each word to a unique ID. The mapping of words with their ID is provided in vocab.txt. And we represent each tweet as a list of (ID, word -count) tuples. Third feature is

Same as “count” except that instead of word counts, we provide the tfidf value as a measure of feature importance. Forth feature is where we mapped each word to a 300-dimensional Glove “embedding vector”.

1. **Literature overview**

The papers written by Rahimi in the paper “Twitter geo-location using Knowledge-Based Methods” (Ref 1) states that their approach outperforms the textual based (where the focus is only on raw data ) .The data set and setting done in this approach was to gather the training data , removing the mentions @ and then using the relations like

1. GEORELATIONS: {isLocatedIn, livesIn, diedIn, happenedIn, wasBornIn }

2. TOP-5 RELATIONS: {isCitizenOf, hasGender, isAffiliatedTo, playsFor, creates }

3. GEO+TOP-5 RELATIONS: Combined GEORELATIONS and TOP-5 RELATIONS. Knowledge base used here is Yago3 and this approach outperformed the other two baseline models namely Proposed Method without Weighted Sum (“simple average”’) and R-GCN baseline method (R-GCN). Their technique uses a graph embedding over the

knowledge base, which they couple with a text representation to learn a geo-location classifier, trained end-to end.

Another paper by **H.T Kung** (Ref 2) presented a data drive approach for twitter geo-location prediction .

The raw tweet is embedded using voca.txt file followed by PCA whitening (technique to speed up unsupervised learning like sparse coding ) in which they used unlabeled data patches to learn basis vectors in dictionary. After this feature extraction technique is applied using sparse coding for geo-location classifier. The approach is based on sparse coding and dictionary learning popular in computer vision for pattern recognization.

1. **Method**

For the experiment , the holdout strategy has been used and hence training were performed on training data sets whereas performance checking was done on dev data sets .

There are three methods which I used for twitter geo-location in which one is baseline classifier and other two are not baselines .

The first classifier I used is zero- base classifier which is simply based on majority class .

On training set of data having raw tweets (train\_full.csv) , I calculated the majority class (the class/region which is the most common among the twitter users.)On getting the majority class , the prediction region will be the same every time .

This is the naïve approach and hence was not useful.

Another method I used is Multinomial Naïve Bayes classifier. For this classifier , the preferred feature type is categorical and hence I have converted the word count feature to all the categorical values .The number of features in each instance is equal to number of unique words in vocab.txt. A binary decision is taken for if a word is present or not by checking the ID and hence 1 or 0 is allotted for that word.

Number of features in each instance equals 2038 and since there are four label class, Multinomial Naive Bayes classifier is used. This approach was pretty good in terms of accuracy and hence is useful.

Another model/classifier which is based on decision tree classifier.

The count vector is used to transform the full data into vectorized format. The accuracy is somewhere better than the baseline classifier. The complete experiment can be found in the README file with complete code . For evaluation metrics , the most important is accuracy which usually defines how accurate the model is .Accuracy is the most important evaluation metrics when it comes to figuring out the model efficacy .Other evaluation metrics are namely precision score , recall score which plays an important role in defining the models efficacy .

Precision is the ratio of the correctly predicted positive observations to the total predicted positive observations. Recall is a metric that quantifies the positive outcome that has made out of all the positive outcomes that could have made.

Recall score tells us about the positive outcomes that are missed.

Out of three classifiers , the best classifier which worked best for me is Multinomial Naïve Bayes .

1. **Results**

The results of three classifiers were good enough to have a look .

The first classifier which is baseline classifier (zero base line) has accuracy less than the other two .

Since the baseline classifier was based on majority class , the accuracy is compromised .The another classifier which works well is decision tree classifier having max depth as 100. Decision tree has first converted the full raw tweet to count vector form and then selecting the best feature out of it using Best feature selection technique and hence creating a model .The accuracy is better than baseline classifier but can be improved . The another classifier which performed better than the two is Multinomial Naïve Bayes classifier and the data used for the prediction is based on categorical values .

This model outperformed the other

models/classifiers .The results are promising enough but can be improved in terms of feature selection .

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Accuracy | Precision | Recall |
| Zero-r baseline | 0.374 | 0.09 | 0.25 |
| Multi-  nomial  Naïve  Bayes | 0.456 | 0.337 | 0.317 |
| Decision tree | 0.391 | 0.230 | 0.261 |

* 1. Evaluation metrics for different classifiers performed.

Above results shows the comparison of different classifiers which predicts geo-location region of twitter users.

The classifier which looks better is the one which has all categorical values and is made from the scratch by me (Multinomial Naïve Bayes) .

The whole idea of the multinomial

Naïve bayes classifier as well as about

other classifiers can be understood

with the code . The classifiers can be

improved to give better results but would take in depth knowledge on

how to find the best features for

geo-location prediction and more .

Results can be seen in the detailed way in the table 1.1.

1. **Discussion**

There are different classifiers which we have discussed here namely zero r

baseline , Multinomial Naïve Bayes classifier and Decision tree .

The features play a critical role in

making a classifier work best .The first classifier is zero r baseline classifier which has two features , first userID and other as raw tweet of the user. Classifier checked for the majority class for training data and the region which belongs to majority of the user is predicted for the dev dataset. As the name suggests , the classifier is based on zero rule . The second classifier which is Multinomial Naive Bayes classifier is based on categorical features. Based on different words in vocab.txt , categorical value is decided to be 1 or 0. If the word is present in the raw tweet then the value is to be 1 otherwise 0. This way all the instances has categorical values .

Predicting a region based on raw tweet can be done by training based on words tweeted by user. But what I feel is there could be better ways to

train the dataset other than using feature like raw tweet. Raw tweets could not be enough to train datasets and hence predict the region of the user. The third classifier which I have worked on is based on decision tree

classifier which is based on count

vector . Having working on raw tweets

of user makes it unethical as reading

personal tweets is not feasible.

There are different sentimental analysis

as well which are done on raw tweets to

check the sentiments of tweet whether

they are good or bad but is unethical.

Getting to work on the private data

raises questions on privacy aspects as well.

1. **Conclusion**

This report consists of details about how region of a twitter user can be predicted using a raw tweet. This also discusses about methods , results about different classifiers .

The methods were mostly text based

But also focuses on some techniques used by others to predict the geo-location easily . Some critical analysis

Is also done in order to understand the problem statement better .

We have discussed different features

used in the process as well. There are three classifiers which I have worked on for geo-location prediction of the user. Different papers are also discussed in which different methods

be it text based or using knowledge based are discussed .The four region mainly NORTHEAST , WEST, MIDWEST and SOUTH.

The operations are to be performed

on the dev data whereas the training

to be done on training data.

1. **References**

Rahimi, Twitter Geolocation using Knowledge-Based Methods

Miriam Cha, Gwon, H. T. Kung, Twitter Geo-location and Regional Classification via Sparse Coding, Proceedings of the Ninth

International AAAI Conference on Web and Social Media 582

1. **Bibliography**

There are two different papers which are mentioned here .First paper which is mentioned is written by students of University of Melbourne in which knowledge based methods are used for location prediction instead of text based .

The method is based on using other relations to determine tweets and can be used to predict regions of tweets. Another paper which is mentioned is written by H.T Kung,

the technique used is sparse coding

which is used for image recognization. Also followed by technique called PCA.