COMP90049 Project 3 Report – Tweet User Geolocation Prediction

1 Introduction

With the advent of social media outbreaking such as Facebook, Instagram and Tweet, which are the three mainstream online networking platforms, the data analysis for users has been a bandwagon both in industries and academia. The "Big Data" has brought many benefits not only to organizations but also for individuals. For example, companies can acquire more customers by popping up their preferred products through email, on website banner or digital application advertising based on their persona analysis. People could also easily access to their favorite things online or offline by the backend data interaction with all directions. The state-of-art behind "Big Data" is the machine learning algorithm.

In this report, the artificial intelligence (AI) for predicting a Tweet user's geolocation will be implemented, through an experiment by three well-known algorithms, Logistic Regression, Neutral Network and Random Forest. The evaluation of model effectiveness will be included, and the knowledge gained from the experiment will also be discussed, from which the best the model among the three will be applied to a real prediction.

2 Literature Review

Previous work with respect to natural language machine learning has reached various achievements. Some popular areas focus on sentiment analysis and geolocation classification for a user. To narrow down, only the geolocation-related stream will be reviewed particularly for this report.

Eisenstein et al. (2010) proposed a multi-level generative model to explore the underlying relationship between tweet topics and geographic areas, which identifying the regional variations of linguistic consistency. The Cascading Topic Model was utilised first to determine a base topic from a chain of random variables (words). Then the Geographic Topic Model was followed and combined for selecting the region topic, regional variance, and the spatial distance, which were used to determine the specific word for a particular location by iterating until a convergence has been reached with the variations in hidden topics.

Pennington et al. (2014) constructed a model to embed each word in a Tweet text with semantical meaning against the local (in text) and global (vocabulary) context, which is called Glove vector, requiring three words must be involved to determine the meaning. A joint model incorporating text feature and user-user interaction in a neural network has been built to form a Multiview Geolocation predictor GCN, with which each layer takes an average of each sample (user) and its immediate neighbors (e.g., friends) by weight for the activation function to infer the user location (Rahimi et al.,2018). Cheng et al. (2010) devised a probabilistic framework to estimate a user's location based purely on the Tweet content with the absence of IP information. The model was finetuned by considering neighbor cities at different level to decide whether the location should be more refined.

3 Methodologies

3.1 Data Pre-processing

The data used in this report refers to the published resource from Eisenstein et al. (2010). Table 1 shows the data structure.

Corpus	Data format for each corpus	Features
Training set	Full tweet text	133,795 instances with label
	Term Frequency (word count)	
	Term Frequency Inverse Document Frequency (tfidf)	
	Embedded semantic meaning (Glove300)	
Development set	Full tweet text	11,475 instances with label
	Term Frequency (word count)	
	Term Frequency Inverse Document Frequency (tfidf)	
	Embedded semantic meaning (Glove300)	
Test set	Full tweet text	12,018 instances without label
	Term Frequency (word count)	
	Term Frequency Inverse Document Frequency (tfidf)	
	Embedded semantic meaning (Glove300)	
Vocabulary library with word ID (2038)		

Table 1 – Dataset for Tweet Geolocation Prediction

The dataset of *Term Frequency* (TF) and *Term Frequency Inverse Document Frequency* (TFIDF) only contains the value of word occurring in the vocabulary library extracted from all tweet messages. In order to achieve the better performance, the entire 2038 vocabulary has been converted to features in *Python*, with all words which are not showing mapped to be zero and all words having corresponding value to the specific feature text.

3.2 Baseline Model

3.2.1 One-R Model

One-R Model is a simple yet comparatively accurate classification algorithm since it requires some feature engineering to select one most predictive attribute before assigning a class from *majority voting* decision rule to an instance. The feature to be chosen should be with the smallest error rate. The following feature selection method in section 3.2.2 has been applied to observe the performance.

Another characteristic for *One-R Model* is that it is equivalently a one-layer *Decision Tree* (DT) which considers the feature interaction so the classifier could be more reliable. Therefore, the *One-R model* has been implemented in *Python* by two ways, manual and decision tree, to observe the different performance and obtain more reasonable benchmark.

3.2.2 Feature Selection

3.2.2.1 Chi-square and Mutual Information (MI)

The idea of the two methods is similar, whereas the determination of *Chi-square* is based on whether the observed value-class combination of a feature occurs more than expected if the feature and class are independent. *MI* considers the sum of every possible value-class combination of a feature weighted by the proportion of instances that actually have that combination (Frermann, 2021).

The two ways has been implemented in Python as well to observe which one could gain the best feature leading to better manual baseline model performance.

3.3 Training Model

3.3.1 Multinomial Naïve Bayes (MNB)

The *Multinomial Naïve Bayes*, a sub-stream of *Naïve Bayes,* is the most suitable model for natural language processing compared to other *Naïve Bayes* classifiers such as *Gaussian* or *Bernoulli* due to the discrete features in the dataset. It is under the *conditional independent assumption* whereas it is still one of the most popular models to classify an instance because of its surprising accuracy.

However, the *MNB* cannot be used on Glove300 training dataset since there are negative float numbers which contradict the data distribution required by *MNB.* Therefore, the model is abandoned since there would be unfair to evaluate the performance over three different dataset types if used.

3.3.2 Multinomial Logistic Regression (LR)

The LR is from a *linear regression* function applied to *sigmoid function* for a prediction that an instance has a class *C*. Each coefficient before X_i represents the association between the feature and class. The *batch optimization* of *gradient descend* is used after each iteration to optimize the coefficient achieving the minimum of *loss function* so the variables X_i could be mostly explanatory. For multinomial prediction, the *softmax function* is used to find the maximum probability of a class over the entire class set to be assigned to an instance.

3.3.3 Multi-Layer Perceptron (MLP)

MLP is a type of feedforward neutral network, which could be an enhanced version of LR, keeping iteratively extracting higher level features from raw data through each layer until the most optimal predictor has been made. Theoretically, it could perform better then LR could, which remains to be proved in this report.

3.3.4 Random Forest (RF)

Random Forest is an advanced version of *D*ecision *Tree.* It implements a combined classification by *majority voting* with the outcomes from multiple training set utilising bagging strategy, which has the minimized the overall variance without interfered by model bias. This model may have the similar performance with *MLP* as *MLP* also aims to achieve the bias and variance trade-off, yet by backpropagation. The two models are powerful enough to achieve the optimal correct prediction.

3.4 Evaluation Metrics

The following metrics are used to compare and evaluate the model performance across the development datasets of three types, from which the best performed model will be selected as the final model applied on test dataset of the corresponding type to make the user geolocation prediction.

3.4.1 Accuracy

Accuracy is the most basic evaluation metric to examine the ratio of correct predicted class over total number of test instances. This is the most intuitive metric to be seen after model application, which gives an intuitional answer regarding the model performance.

3.4.2 Precision

Precision indicates the rate of how many correct predictions has been made among all the predicted classes interested, which should be as high as possible.

3.4.3 Recall

Recall reveals the rate of how many correct predictions has been made among all the true labels interested, which should also be as high as possible to decrease the false negative rate.

3.4.4 F1-Score

F1-Score is the balance version of Precision and Recall, expected to be high.

4 Results

4.1 Feature Selection

As can be seen from the Table 2, Chi-square and MI selected different best feature for the baseline model. From the training dataset of word count (D^{Ttf}), exclamation mark was selected as the most predictive text for a geographical classification, whereas MI chose “”. By comparison, from the training dataset of TFIDF (D^{Ttfidf}), "@user_559b1bbb" and "ima" were chosen respectively. Glove300 (D^{T300}) could not be applied for Chi-square due to the negative float number, but MI picked “bad” for the best feature.

train_count

One-R	Chi-square		MI	
	Word ID (0)	Accuracy	Word ID (946)	Accuracy
Manual	"!"	0.37	"isnt"	0.37
Decision Tree	-	0.41	-	0.41

train_tdidf

One-R	Chi-square		MI	
	Word ID (144)	Accuracy	Word ID (929)	Accuracy
Manual	"@user_559b1bbb"	0.12	"ima"	0.37
Decision Tree	-	0.41	-	0.41

train_glove300

One-R	Chi-square		MI	
	Word ID	Accuracy	Word ID (254)	Accuracy
Manual	-	-	"bad"	0.37
Decision Tree	-	0.41	-	0.41

Table 2 – Feature Selection Analysis

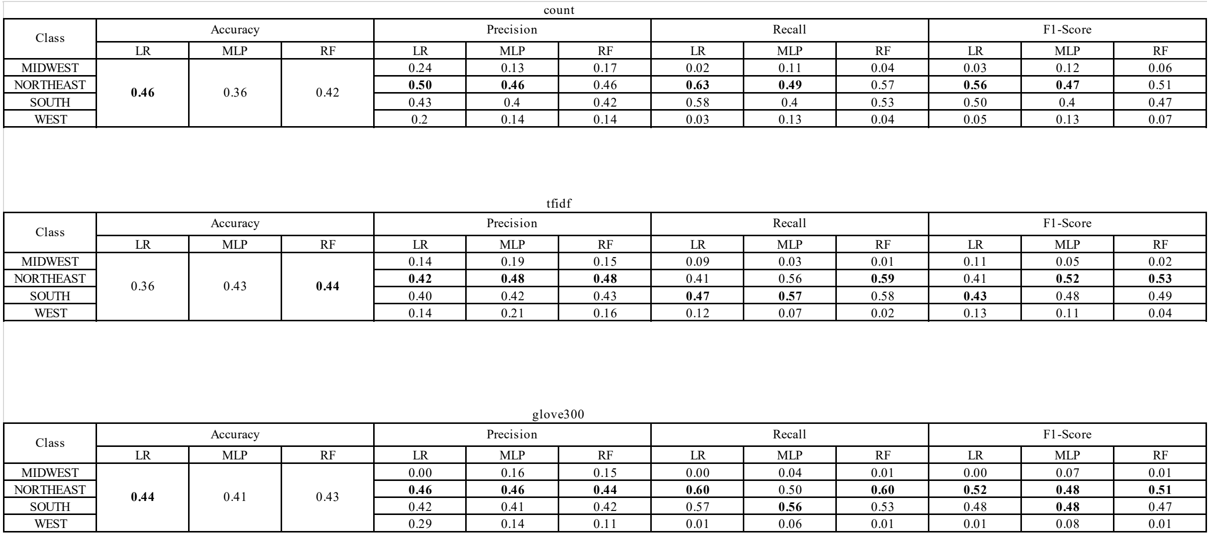
Some tendency that the word selected were being more and more semantic can be concluded from this. From D^{Ttf}, TF is the main value of each feature, and “!” appears massive times in Tweet texts so it was picked based on Chi-square mechanism, nevertheless it is only a punctuation without real meaning as if a word. As feature selection went with different feature value type, the word with more semantic meaning was picked. The reason why this happened was that rather than simply counting words, TFIDF ensures a word has enough frequency in a text (to be specific enough) while has not too much frequency for all texts (to be not too general), which produces a trade-off to ensure the word is predictive enough from a text so the word meaning starts to emerge. On the other hand, Glove embeds the meaning to a word by assessing a word itself with all other words in the texts against the whole words around a particular word it is compared to (Pennington et al., 2014). In this way the word picked from the text would have great meaning for predicting a region class.

By using these features to manually train One-R Model, the accuracy is still not good because “NORTHEAST” distributes dominantly in region labels, which leading to massive wrong prediction based on *majority voting*.

On the flip side of the coin, one-layer DT has a reasonable performance, with 41% correctness, which corresponds to the characteristic of DT.

4.2 Model Performance

The overall model performance was not ideal. By scrutinizing the metrics, some reasons may be identified as follows. With such big data pool, the number of layers of MLP is only three, which may lead to overfitting so the performance of MLP was not ideal as expected. However, the Logistic Regression is unexpectedly underperformed on the second dataset, which could resulted by feature value variation, whereas the accuracy of Logistic Regression outperformed on two datasets, and with fairly high Precision, Recall and F1-Score on D^{Ttf}, this model was selected as the final model for real prediction applying on the dataset valued by TF.



count

Class	Accuracy			Precision			Recall			F1-Score		
	LR	MLP	RF	LR	MLP	RF	LR	MLP	RF	LR	MLP	RF
MIDWEST	**0.46**	0.36	0.42	0.24	0.13	0.17	0.02	0.11	0.04	0.03	0.12	0.06
NORTHEAST				**0.50**	**0.46**	0.46	**0.63**	**0.49**	0.57	**0.56**	**0.47**	0.51
SOUTH				0.43	0.4	0.42	0.58	0.4	0.53	0.50	0.4	0.47
WEST				0.2	0.14	0.14	0.03	0.13	0.04	0.05	0.13	0.07

tfidf

Class	Accuracy			Precision			Recall			F1-Score		
	LR	MLP	RF	LR	MLP	RF	LR	MLP	RF	LR	MLP	RF
MIDWEST	0.36	0.43	**0.44**	0.14	0.19	0.15	0.09	0.03	0.01	0.11	0.05	0.02
NORTHEAST				**0.42**	**0.48**	**0.48**	0.41	0.56	**0.59**	0.41	**0.52**	**0.53**
SOUTH				0.40	0.42	0.43	**0.47**	**0.57**	0.58	**0.43**	0.48	0.49
WEST				0.14	0.21	0.16	0.12	0.07	0.02	0.13	0.11	0.04

glove300

Class	Accuracy			Precision			Recall			F1-Score		
	LR	MLP	RF	LR	MLP	RF	LR	MLP	RF	LR	MLP	RF
MIDWEST	**0.44**	0.41	0.43	0.00	0.16	0.15	0.00	0.04	0.01	0.00	0.07	0.01
NORTHEAST				**0.46**	**0.46**	**0.44**	**0.60**	0.50	**0.60**	**0.52**	**0.48**	**0.51**
SOUTH				0.42	0.41	0.42	0.57	**0.56**	0.53	0.48	**0.48**	0.47
WEST				0.29	0.14	0.11	0.01	0.06	0.01	0.01	0.08	0.01

Table 3 – Model Performance Analysis

5 Discussion and Improvement

Based on the performance observed, the overall training process could be improved. Firstly, the *cross-validation* strategy should be applied to ensure the model is of low variance and not to be overfitting or underfitting. The development dataset could be combined with training dataset to form a whole dataset for different cross-validation training so the model effectiveness could be improved enormously. Second, MLP is sensitive to different scale of input value so the scaler could be used for different data input. Also, different depth of hidden layer and width of each layer could be experimented for exploring the better model performance. Pipeline should also be used to avoid data leakage when applying *cross-validation*. Third, various dimension of feature could be chosen for feature selection, then applying different model could gain different results.

6 Conclusion

Overall, the Tweet Geolocation Prediction were experimented as an initial foray. With coding and thinking implemented, numerous improvement could be applied in the next stage for making progress about studying machine learning.

7 References

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287.

Frermann. L. (2021), Tutorial Material, Topic: "Model Evaluation, Feature Selection and Analysis", School of Computing and Information Systems, Melbourne University, Melbourne, Vic., Apr., 2021.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rahimi, A., Cohn, T., and Baldwin, T. (2018). Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2019.

Schu ̈tze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.