# COMP90049 - Assignment 3 Report

## 1 Introduction

The boost of fast-consuming social media like Twitter has swept over the world and construct a significant proportion of online data. Mining real-time and sparse data like this has a promising influence on promoting personalised advertisement, analysis of social science and novel geographical-based applications (earthquake detection by Twitter (Sakaki, Okazaki and Matsuo, 2010)). Unfortunately, only few users have tagged their content with geo-related information such as country, city, along with some non-sensical locations like "Wonderland", making the geographical prediction through raw text much more difficult than general classification problems. Moreover, Twitter users tend to use shorthand and informal vocabularies for communication, addressing the importance of feature engineering on tweets content.

The experiment has made use of a pre-processed dataset (Eisenstein et al., 2010), namely a set of tf-idf value of tweets text to predict one of four locations (MIDWEST, NORTHEAST, SOUTH, WEST) of a user. Two datasets are used: train and development set (Eisenstein et al., 2010), with the former one for training and the latter for testing. Two classification algorithms were attempted: Multi-Layer Perceptron (referred to **MLP**) and Multinominal Naïve Bayes (referred to **MNB**). 0R baseline was operated to compare the performance of classifiers. The proposed approach on finding the optimal classifiers is trail-and-error to find the hyperparameters that lead to the optimal performances.

## 2 Literature review

Eisenstein et al. (2010) have proposed a model that processes geo-classification by analysing lexical variation on topics of raw text which includes insights about regional slangs. The main dataset used was gathered from Twitter official API that is the same as this experiment applied. The main assumption of their model is that regions and topics have high correlation on shape observed lexical frequencies. They have combined Geographical Topic Model with Cascading Topic Models that generates text from random variables to indicate topic variations, which produce the latent variable *r* corresponds to the geographical region of the author. There are many variational distribu tions applied on latent variables to form this hierarchical generative model. As a result, it outperformed two baseline models, each with their optimal results.

Cheng et al. (2010) has proposed a probabilistic framework that predict users' city-level location based on purely tweets text. They determined the spatial focus and dispersion of words in tweets and then evaluate maximum likelihood with the occurrence of words as well as distance between a city and the model's centre. The model relies on two refinements: (i) identified words in tweets with a strong geo-scope; and (ii) a lattice-based smoothing model. Their prediction result is 51% which is reasonably good.

Considering approximately only 1% of geotagged tweets are applied on supervision, Rahimi et al. (2018) has proposed a transductive multiview geolocation model on a semi-supervision level, using Graph Convolutional Networks (Kipf and Welling, 2017) which outperforms two strong baselines (MLP and DCCA (Andrew et al., 2013)). However, when given sufficient supervision, MLP produces better results than others. Under semi-supervised occasion, GCN exceeds the text-only, network-only, and hybrid geolocation models when using both text and geo-related social network to infer users' location in a joint setting.

## 3 Method

Without feature engineering process, this experiment has applied tf-idf dataset as input, where it transformed raw text into a vector that contains each term with its corresponding composite weight (Manning et al., 2018). The tf-value has utilized the frequency of each word in the dictionary file where it should be zero when the word has not appeared in a tweet. Hence, each instance in the dataset is a vector with a size of total number of words in dictionary.

While taking the advantages of tf-idf value, choosing the right classifier is a crucial step. Although intuitively, applying KNN and Logistic Regression are more suitable for these continuous values, time consumption is also an important metric in terms of training cost; thus, after trying different models, it is obvious

that most of them cannot even achieve better than MNB (with most accuracies are less than 0.5). After random attempts, the experiment focused on MNB and MLP to find the optimal hyperparameters that lead to their best outcomes.

The evaluation metrics used are accuracy and F1-score for all classifiers. Accuracy is more emphasized and being tested around. One baseline model (0R baseline) was performed to compare the performances of other classifiers. The method used for finding the optimal result is trial-and-error by tuning hyperparameters. Main hyperparameter for MNB is alpha which is additive Laplace smoothing parameter whose value is ranging from 0.1 to 1, while MLP has three hyperparameters tested: max-iteration, learning rate and activation function for hidden layers. MLP is time-consuming; therefore, it is taken a deeper look into changing max-iteration to see when the model reaches its convergence and their corresponding performances while modifying other two hyperparameters.

## 4 Results

0R baseline model has 0.374 accuracy and 0.136 F1-score. While using MNB, alpha is changing from 0.1 to 1. Through table 1, it is obvious that both the accuracy and F1-score have not fluctuated greatly as alpha went higher. Accuracy and F1-score are approximately 0.464 and 0.27 respectively. The results are very steady and performed better compared with the baseline.

| alpha | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 0.463 | 0.463 | 0.463 | 0.463 | 0.464 | 0.464 | 0.464 | 0.464 | 0.464 |
| F1-score | 0.27 | 0.27 | 0.269 | 0.269 | 0.269 | 0.269 | 0.269 | 0.269 | 0.269 |

**Table 1** – MNB performances with alpha in (0,1)

To have an insight on how MNB performs on training and development datasets, figure 1 is used to compare their accuracies. Same situation arises when we plot them on training set which stick around 0.49. This value also indicates that MNB was not overfitting the training set.

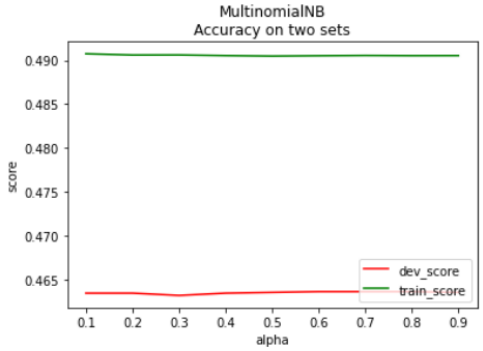


**Figure 1** – MNB accuracy (alpha in range (0.1,1))

There are three hyperparameters tuned in MLP: max-iteration, activation function and learning rate. Figure 2 has shown the accuracies on different settings. When activation is logistic and learning rate is constant, the classifier converges when max-iteration is greater than 35 (as warning messages) but it converges from 30 when learning rate is adaptive. This occasion is also indicated by stable accuracies on figure 2 when activation equals to logistic with most accuracies are under 0.46. Overall, logistic function performs better than relu; under which circumstance the accuracy reaches its highest as 0.46 when max-iteration is 60.

While accuracies for adaptive and constant activation functions are close to each other, F1-score has a different story. From figure 3, relu presents better results although it goes down when max-iteration exceeds 50. The scores fluctuate a lot as max-iteration change.

Considering accuracies of these models, further range of max-iteration was extended to (1,30) as the convergence appears around 30 to 35 for logistic activation function (table 2 and 3 only shows data in range (1,13) for simplification).

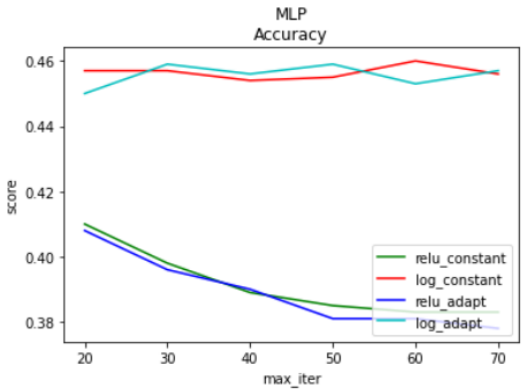


**Figure 2** – Accuracy with max-iteration in (20,70)

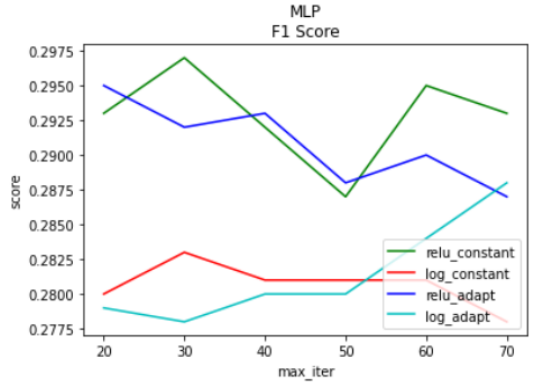


**Figure 3** – F1-score with max-iteration in (20,70)

As table 2 indicates, the highest accuracy after training is 0.465 when using logistic function and adaptive learning rate. Figure 4 also implies that when using relu function, accuracy decreases as it grows. In general, F1-score (See table 3) is increasing in range (1,13). Although the models have not reached convergence with small max-iteration values, they performed well in terms of accuracy.

By comparing the results in two ranges, logistic function often performs better than relu, while adaptive learning rate usually outperforms constant in both accuracy and F1-score. Therefore, final hyperparameters settings for MLP are max-iteration equals 3, activation is logistic and learning rate is adaptive.

Comparing best classifiers of MNB and MLP, MLP wins with higher accuracy (0.465) with approximately similar F1-score (0.268) with MNB. Hence, the final classifier chosen for prediction is MLP.

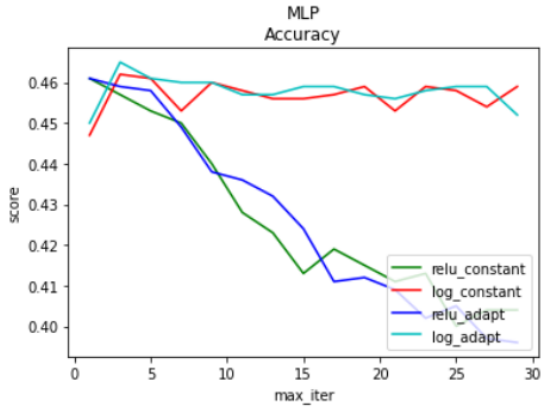


**Figure 4** – Accuracy with max-iteration in (1,30)

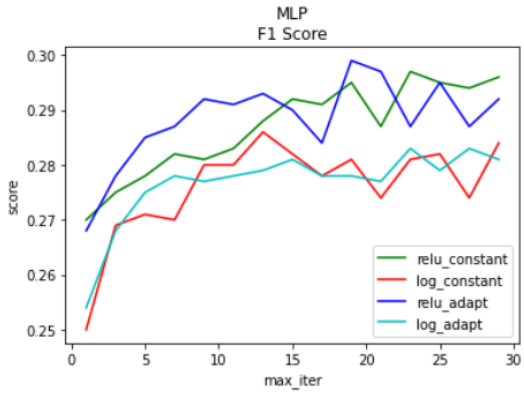


**Figure 5** – F1-score with max-iteration in (1,30)

| max iteration | logistic, constant | logistic, adaptive | relu, constant | relu, adaptive |
|---|---|---|---|---|
| 1 | 0.447 | 0.45 | 0.461 | 0.461 |
| 3 | 0.462 | 0.465 | 0.457 | 0.459 |
| 5 | 0.461 | 0.461 | 0.453 | 0.458 |
| 7 | 0.453 | 0.46 | 0.45 | 0.449 |
| 9 | 0.46 | 0.46 | 0.44 | 0.438 |
| 11 | 0.458 | 0.457 | 0.428 | 0.436 |
| 13 | 0.456 | 0.457 | 0.423 | 0.432 |

**Table 2** – MLP accuracy with max-iteration in (1,13)

| max iteration | logistic, constant | logistic, adaptive | relu, constant | relu, adaptive |
|---|---|---|---|---|
| 1 | 0.25 | 0.254 | 0.27 | 0.268 |
| 3 | 0.269 | 0.268 | 0.275 | 0.278 |
| 5 | 0.271 | 0.275 | 0.278 | 0.285 |
| 7 | 0.27 | 0.278 | 0.282 | 0.287 |
| 9 | 0.28 | 0.277 | 0.281 | 0.292 |
| 11 | 0.28 | 0.278 | 0.283 | 0.291 |
| 13 | 0.286 | 0.279 | 0.288 | 0.293 |

**Table 3** – MLP F1-score with max-iteration in (1,13)

# 5 Discussion

## 5.1 Critical Analysis

Naïve bayes is the most common classification model applied in real life. It calculates the posterior probability of a label by giving the prior probability with conditional probability of features. Generally, each word is conditionally independent with each other. MNB implements algorithm for multinomially distributed data. Intuitively, a label of a tweet can have a multinominal correlation to the words with the whole distribution parameterized by a smoothed maximum likelihood vector $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$, i.e., relative frequency counting. Although using word count seems reasonable in MNB, tf-idf value which has taken the word frequency into account is more accurate which is standardised in numeric values, rather than simply counting each word.

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Alpha is a part of smoothed maximum likelihood (equation above) with a value of 1 meaning Laplace smoothing. The conducted experiment on MNB has indicated that it is very stable both in accuracy (0.46) and F1-score (0.27); thus, having a good performance on text classification for locations.

With the former research on MLP performing text classification (Do et al., 2017), MLP is

another good method which considers non-linear and more complex relationship between features and labels. MLP is sensitive to feature scaling, and tf-idf value fits this requirement properly. We aim to reach the convergence of MLP, suggesting the convergence of loss function (Cross-Entropy). The hyperparameter - hidden layer sizes are set as default, which means the number of layers is 1 with 100 neurons per layer. With 1 hidden layer, it produces a result like Logistic Regression which minimise error of prediction; however, MLP has an advanced method which is back propagation to update the parameters for each word appeared in a tweet. The experiment focused on the activation function and learning rate. However, as max-iteration increases, the accuracy falls. This is due to overfitting on the training set, leading to low accuracy and F1-score. As a result, the best max-iteration number after experiments is 3. Adaptive learning rate concludes that it is adaptive to the performance of the model if it has not improved.

Interestingly, the best result of MNB (accuracy=0.464, F1-score=0.269) and MLP (accuracy=0.465, F1-score=0.268) are very close to each other.

### 5.2 Ethical Issues

The linguistic difference on tweets text impacts a lot on geo-classification problem, which results in bias in terms of demographic differences. In a general research on multiple geo-locating classification dataset, it is phenomenal that women tend to use less geo-related terms, while young people use significantly more non-standard words in a tweet (Pavalanathan and Eisenstein, 2015). This introduces a bias such that when it captures more non-standard words in a tweet, it directs the classification towards a label like that of young people. Moreover, as input data has not incorporated the age of users, it has high possibility to assign labels like young people which takes up a large proportion among Twitter users (Pavalanathan and Eisenstein, 2015).

Geographical bias also exists in learning process. As the degree of urbanization and demographic characteristics vary a lot (Pavalanathan and Eisenstein, 2015), the influence of tweets containing geo-related words has been enlarged during classification. The urban biases lean towards some representative of more tagged areas than less common ones.

## 6 Conclusion

Former researchers have studied geo-locating problem in terms of lexical variables (Eisenstein et al., 2010), spatial dispersion of words (Cheng et al,. 2010) and GCN for semi-supervised learning (Kipf and Welling, 2017). This paper has focused on two model algorithms: MLP and Naïve Bayes, which are easy to implement by applying tf-idf value of tweets. As tf-idf featuring the importance of each word as continuous values, it is aligned with the requirement of MLP properly. By multiple attempts of tuning hyperparameters, they reached their optimal results with similar accuracies (MNB is 0.464 and MLP is 0.465) and F1-scores (MNB is 0.269 and MLP is 0.268). We prefer higher accuracy value; thereby, applying MNB on the final test dataset.

Ethical issues emerge on both demographical and geographical levels because of tweets content. Many studies have emphasized on linguistic relation between tweet content and geo-related terms as well as the impact of social interaction on network. Some models have taken the network into consideration but ignored the linguist variation (Kipf and Welling, 2017). In addition, many useless punctuations create significant noise of data. In the future work, lexical analysis will be emphasised, being applied on both model logics and feature engineering.

## References


Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users. *Proceedings of the 19th international conference on World wide web*. https://doi.org/10.1145/1772690.1772777

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

Manning, C., Raghavan, P., & Schütze, H. (2018). *Introduction to information retrieval* (pp. 117-119). Cambridge University Press.

Rahimi, A., Cohn, T., & Baldwin, T. (2018). Semi-supervised user geolocation via graph convolutional networks. *arXiv preprint arXiv:1804.08049*.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning* (pp. 1247-1255). PMLR.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet. *Proceedings of the 19th ACM international conference on information and knowledge.* https://doi.org/10.1145/1871437.1871535

Do, T. H., Nguyen, D. M., Tsiligianni, E., Cornelis, B., & Deligiannis, N. (2017). Multiview deep learning for predicting twitter users' location. *arXiv preprint arXiv:1712.08091*.

Pavalanathan, U., & Eisenstein, J. (2015). Confounds and consequences in geotagged Twitter data. *arXiv preprint arXiv:1506.02275*.

Eisenstein, J., Smith, N. A., & Xing, E. (2011, June). Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 1365-1374).