# COMP90049 Project 2: Geolocation of Tweets with Machine Learning

**Anonymous**

## 1 Introduction

Twitter is a well-known social media and users can send tweets to share their opinions with others online. It is evident that tweets sent from users in different locations normally have different features. By analyzing the features of tweets sent by different users, we can predict the location of twitter users.

In this project, I will make predictions on the location of some twitter users based on the analysis of tweet features and the authors' locations. This report aims to use machine learning methods to geolocate different tweets.

## 2 Literature Review

Some related works have shown that the KNN algorithm is a great machine learning algorithm to solve classification problems. In [1], the author has done an experiment and concluded that KNN could help classify some particular data. Thirunavukkarasu, K et al. have used KNN as the machine learning algorithm and the result of evaluation shows that the machine learning model is quite effective [2]. These works are related to KNN algorithms usage in classifying data. In this project, the problem is also classifying tweet data into three different classes, which are the predicted location of the tweets. Thus, KNN is a suitable algorithm to use.

## 3 Dataset

The data used in this project are made up of many different tweets sent from different twitter users, which originate from [3-4]. These tweets are divided into three different parts, namely training data, developing data and testing data. The training data is used to train the machine learning model. The developing data is used to evaluate the effectiveness of the model that has already been trained. The testing data is used to generate the location of twitter users based on the given data.

The dataset has already been preprocessed so that the task becomes simplified. The term frequencies have been recorded and stored in separate files based on different standards. In this project, these files will be used to train the machine learning model as well as evaluate and test the model.

The locations of twitter users have been limited to three states, namely New York, California and Georgia so that the task can be easily solved.

## 4 Evaluation Metrics

In this project, I will use the following metrics to evaluate the effectiveness of geolocating different tweets: precision and recall. Precision refers to the percentage of correct predictions among all the predicted positive instances. Recall refers to the proportion of correctly identified instances among all the real positive instances [5]. In this project, precision and recall are quite different in different classes as it is a multi-class problem. Therefore, I will use Micro-Averaging method to calculate the overall precision and recall of the whole system:

$$\text{precision}_M = \frac{\sum_{i=1}^{3} true\ positives(i)}{\sum_{i=1}^{3} true\ positives(i) + \sum_{i=1}^{3} false\ positives(i)};$$

$$\text{recall}_M = \frac{\sum_{i=1}^{3} true\ positives(i)}{\sum_{i=1}^{3} true\ positives(i) + \sum_{i=1}^{3} false\ negatives(i)}.$$

In this formula, true positives(i) refers to the true positives value of i-th class; false positives(i) refers to the false positives value of i-th class; false negatives(i) refers to the false negatives value of i-th class [5]. In this project, we assume that the first class is New York, the second class is California and the third class is Georgia.

Typically, models with high precision

have low recall and vice versa. To make the evaluation effective, I will also use F-score. In this project, I use F-1 only to give the same weight for precision and recall in the evaluation process. The formula of F-1 is shown below:

$\mathrm{F}_1 = \frac{2*precion*recall}{precion+recall}$. If the F-1 value of the model is high, it means the model is effective. In this project, I will use the weighted average F-1 score as the overall F-1 score for the three classes.

# 5 Methodology and Result

## 5.1 Methodology

The K-Nearest Neighbor (KNN) algorithm is used in this project. KNN is an effective method in classifying. This method is based on the distances between the test input and other training instances. In this algorithm, the output class is decided by the majority class of k nearest training instances. The reason for using this method is that tweets sent from different locations normally have very similar features and that tweets sent from different locations often have quite different features. So, it is suitable to use the KNN algorithm to geolocate different tweets.

I have chosen K=1, 3, 5 and 7 in the KNN algorithm and compare the effectiveness of geolocating tweets among different K values. The training data and testing data for all trails are the same so that the only difference in the comparison is the K value.

In this project, the best-200 file is used as the preprocessing result of the raw tweet data. This is because the terms with the greatest Mutual Information and Chi-Square values are very useful for the model to make predictions on the location of different tweets. Apart from that, top 200 terms can best represent all the features. Therefore, I choose best-200 file instead of other given files. Specifically, train-best200.arff is used to train the model and dev-best200.arff is used to evaluate it.

In the experiment, I have used Weka as a tool to run the machine learning algorithms based on the input training data and evaluation data. After generating the output of each given instances, Weka will automatically evaluate the effectiveness of the model that has been selected. Therefore, Weka is an efficient tool to train and evaluate machine learning model.

## 5.2 Result

The result of using KNN algorithms to evaluate the data is presented in the table below.

| Value of K | $precision_M$ | $recall_M$ | $F_{1_M}$ |
|---|---|---|---|
| 1 | 58.1% | 60.7% | 58.8% |
| 3 | 62.7% | 66.1% | 60.8% |
| 5 | 65.2% | 66.9% | 59.3% |
| 7 | 62.6% | 65.1% | 55.4% |

Table 1: result of experiments using KNN

From the table, we can see that the precision of the 5-NN model is the greatest among four KNN models. Meanwhile, 5-NN also has the greatest recall value. However, 3-NN has the greatest F-1 score. In terms of 1-NN, the precision and recall of it are incredibly low, much lower than that of other models. Therefore, 3-NN performs the best in regard to geolocating different tweets.

# 6 Critical Analysis

The algorithm has done well to geolocate tweets, and precision of the algorithm has achieved around 60%, which is not bad. However, the precision value can be improved further so that the geolocation of tweets can be done more precisely. In the raw tweet data, there are many typos and abbreviations which are the same as the original term, like “talkin” is the same as “talking” and “u” is the same as “you”. These terms should have been considered the same when we preprocess the raw data, so that the precision of the model can be enhanced.

In terms of the KNN algorithm itself, it is a pretty good algorithm. Tweets from different locations normally have very different features so the distances between them should be a good way to classify different tweets. However, different terms should not have equal weight when calculating the distances between instances. In my opinion, more weight should be given to terms that are nouns, verbs and adjectives. The subject and the object of a given tweet should have more weight than other conjunctives. For example, in the tweet “I'm the spokes person for extenze”, the term “I’m”, “spokes”, “person” and “extenze” should be given more weight and “the” and “for” should be given less weight.

## 7 Conclusions

After using the KNN algorithm, the precision and recall of geolocating tweets are satisfactory. However, if we take some measures to enhance the machine learning algorithm, the result of geolocating tweets can become better.

## References

[1] Qu Song. The comparison and analysis of classification methods for psychological assessment data. The 2nd International Conference on Information Science and Engineering, IEEE, 2010.

[2] Thirunavukkarasu K, Singh Ajay S, Rai Prakhar,Gupta Sachin. Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning. 2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE, 2018.

[3] 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.

[4] Rahimi, Afshin, Trevor Cohn, and Timothy Baldwin. Semi-supervised user geolocation via graph convolutional networks. arXiv preprint arXiv:1804.08049 (2018).

[5] Renuka Joshi et al. Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/, 2016.