Paper template for COMP30027 Report

Anonymous

Abstract

1. Introduction

This project’s aim is to use various classifiers to identify the location of a textual message (a tweet specifically) and analyze those classifiers. Random forest, logistic regression, and deep learning are the classifiers that have been chosen for this project. Short message location identification can be perceived as a simplification of geottaing. By definition, geotagging is an action of attaching a geotag is “a piece of electronic data that shows where someone or something is [...]” ("Meaning of geotag", n.d.).

2. Related Works

*Machine learning (*ML) has been used to categorize texts into desired categories has been around for years already (Sebastiani, 2001). Within this field, there is a more specific subsection which revolves around identifying a text to its origin, a physical location on Earth. The presence of books about this specific matter highlights the matter’s popularity. An example of a book that was devoted to identification of a message’s location is *A Machine Learning Approach for Resolving Place References in Text.* by Bruno Martins, Ivo Anastácio, and Pável Calado ().

When it comes to methodology, there have been various attempts. One of the studies used Support Vector Machines (Avvenuti & Cresci, 2018). Others created their own methodology, such as Martins, B., Anastácio, I., & Calado, P. (2010) who have developed a new ML method “for resolving place references in text, i.e. linking character strings in documents to locations on the surface of the Earth”. Hence, basing on the years those studies were published, one can conclude that this field is rather new. It could be possibly due to “the falling cost of large data storage devices and the increasing ease of collecting data over networks□”, “the development of robust and efficient machine learning algorithms to process this data”, and “the falling cost of computational power□ enabling the use of computationally intensive methods for data analysis” ()

3. Data and Classifier Description

In this section, the usage of data and chosen machine learning methods will be described. First, preparation of data will be discussed to show how the datasets were groomed in order to be provided to the methods. Then, each method will be defined.

3.1. Preparation of data

For our research we used the prepared data that was provided on the lms. It was processed from the raw tweet texts in the following way. First the punctuation and capitalisation of letters were removed. After this the full vocabulary of all the tweets is determined. Of all these words the ones who are most “present” in the tweets of a certain city, according to Chi square (cite) and Mutual Information Criteria (Yang & Pedersen, 1997), are selected as features in a bag of words model (cite). The best words of every city are combined while removing duplicates to form the final feature set. The frequencies of all featured words are calculated per tweet and stored to create the final dataset.

This method seems perfectly fine to work with, however there would have been a lot of other ways to process this data. We also looked at removing stopwords from tweets and the lemmatization of words (using a dictionary to categorize small differences in words under the same word) (cite) as but our implementations did not seem to improve the accuracy of our classifiers. This is in line with the findings of Saif et al (2014) and Bao et al (2014)

Embedding was attempted for this work, but the processing time was so long (due to several thousands of variables) that this was not deemed worthy of pursuing. Doing embedding, understanding its meaning, tuning it, and then interpreting it with significance would take weeks to complete.

3.2. Random Forest Classifier

"Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (Breiman, 2001). The classifier is an ensemble of Random Trees, which according to Nicholson, Baldwin, & Verspor, (2019) are defined as a "decision tree [...] where at each node, only some of the possible attributes are considered". Ensemble methods are "learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions" (Dietterich, 2009).

3.2. Logistic Regression Classifier

Logistic Regression (LR) is a model that adapts linear regression in a way such that it can be used as a classifier. It reaches this result by incorporating the logistic function. It can be fitted on a dataset by utilizing maximum likelihood estimation. Because of it's simplicity and interpretability, Logistic regression is a staple classification method for statisticians (Genkins, 2007).

3.3. Stacked Ensemble System

There are two parts that describe this system: stacking and ensemble. In the process of stacking "an algorithm takes the outputs of sub-models as input and attempts to learn how to best combine the input predictions to make a better output prediction" (Brownlee, 2018). Hence, a stacked ensemble system is supervised ML algorithm that ensemble of chosen algorithms and uses stacking.

Güneş, Wolfinger, & Tan (2017) define Stacked Ensemble System (SES) as "combin[ation] [of] predictions from multiple machine learning algorithms and use these predictions as inputs to second-level learning models." Meta-classifier is another way to call second-level learning model (Nicholson, Baldwin, & Verspor, 2019).

For this work, Decision Tree and Gaussian Naive Bayes Classifiers were chosen as the base (first-level) learning models. Their outputs have been provided to the meta-classifier, Logistic Regression.

3.3.1. Decision Tree Classifier

Pal, & Mather (2003) describe Decision Tree (DT) as "computationally fast, make no statistical assumptions, and can handle data that are represented on different measurement scales".

3.3.2. Gaussian Naive Bayes

Gaussian Naive Bayes is defined as an "extension of Naive Bayes". It uses "Gaussian (or Normal distribution) [which] is the easiest to work with because you only need to estimate the mean and the standard deviation from your training data." (Brownlee, 2016).

3.3.3. Logistic Regression

A definition of Logistic Regression have been provided in the previous subsection.

4. Evaluation of the Classifiers

Table 1 below showcases the accuracies of each method that have been used for this work. It is not a surprise that ensemble methods score higher than a single standing classifier (Logistic Regression in this case). This argument is supported by Nicholson, Baldwin, & Verspor, (2019), because they argue that "the combination of lots of weak classifiers can be at least as good as one strong classifier". Ensemble allows for voting and since base classifiers are independent, they hardly ever make the same mistake. Hence, it is hard for majority of classifiers to choose the wrong answer. This can be proven mathematically too (Nicholson, Baldwin, & Verspor, 2019).

Method	Test Accuracy
Random Forest	29.882%
Logistic Regression	30.027%
Stacked Ensemble System	30.041%

Table 1: Methods' accuracies

SES has a higher accuracy than RF. This can be due to the fact that SES has a more varied base classifiers. In this work, SES had DT and LR as the base classifiers, and RF has only multiple DT. The variety of base classifiers

Since the output is of categorical data type, standard deviation and variance cannot be calculated because they can be determined only if the output is numerical, where there is order among instances and difference between each instance is calculable. However, recall and precision can be computed. Table 2 and Table 3 respectively shows the recall and precision of each ML method.

Saxena (2018) says that precision can be calculated by dividing true positives over the sum of true positives and false positives. Recall, though is calculated by dividing true positives by the sum of true positives and false negatives (Saxena, 2018). She also outlines the difference between them: “[p]recision means the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by your algorithm”. Thus,

Method	Recall
Random Forest	29.783%
Logistic Regression	30.027%
Stacked Ensemble System	30.041%

Table 2: Methods’ recalls

Method	Precision
Random Forest	51.918%
Logistic Regression	50.345%
Stacked Ensemble System	52.004%

Table 3: Methods’ precisions

Precision and recall have an inverse relationship — if we want to improve one of them, the other one will usually get worse (Nicholson, Baldwin, & Verspor, 2019), which seems to be the case of all classifiers. An interpretation of those results could be that on average “the proportion of the data points our model says was relevant actually were relevant” was bigger the proportion of the data points being actually irrelevant. However, the precision is so close to 50% that it is essentially similar to chance, like when one flips a coin. Hence, the methods’ precision have similar success rate as randomly flipping a coin, which are not good news.

Another reason for such poor scores could be the fact that short messages that were used for the datasets were from Twitter. The tweets tend to contain a lot of words that often are modified as part of the colloquialism. For example, “good” could be written as “goooood” to highlight this word. The machine does not know that those are the same and treats as different words, which can impair how good the algorithm scores.

5. Conclusions

This work has shown that short message identification is not an easy task. The accuracies are far from acceptable and the precision as well as recall are not great either. This means that there is still a lot work that has to be done to successfully and with confidence identify the geographical origin of the message. Colloquialism is only the tip of the hardships that algorithms have to overcome.

References

Avvenuti, M., & Cresci, S. (2018, June 3). GSP (Geo-Semantic-Parsing): *Geoparsing and Geotagging with Machine Learning on Top of Linked Data*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-93417-4_2.

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014, August). The role of pre-processing in twitter sentiment analysis. In International

Breiman, L. (2001, October). Random Forests. Retrieved from https://link.springer.com/article/10.1023/A:1010933404324

Brownlee, J. (2016, April 11). Naive Bayes for Machine Learning. Retrieved from https://machinelearningmastery.com/naive-bayes-for-machine-learning/.

Brownlee, J. (2018, December 13). How to Develop a Stacking Ensemble for Deep Learning Neural Networks in Python With Keras. Retrieved from https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/.

Dietterich, D. (2000, December 1). Ensemble Methods in Machine Learning. Retrieved from https://link.springer.com/chapter/10.1007/3-540-45014-9_1.

Donges, N. (2018, February 23). The Random Forest Algorithm. Retrieved from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd.

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. Technometrics, 49(3), 291-304

Güneş, F., Wolfinger, R., & Tan, P. (2017). *Stacked Ensemble*

Models for Improved Prediction Accuracy. Retrieved from https://pdfs.semanticscholar.org/43e7/de1ff5fec94f9859727010d0c5b80190286d.pdf.

Martins, B., Anastácio, I., & Calado, P. (2010, March 31). *A Machine Learning Approach for Resolving Place References in Text*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-12326-9_12.

Meaning of geotag in English. (n.d.). Retrieved from https://dictionary.cambridge.org/dictionary/english/geotag.

Mitchell, T. (1999 November). *Machine Learning and Data Mining. 42*(11) 1-1. Retrieved from https://www.ri.cmu.edu/pub_files/pub1/mitchell_tom_1999_1/mitchell_tom_1999_1.pdf.

Nicholson, J., Baldwin, T. & Verspoor, K. (2019). *COMP30027 Machine Learning Evaluation II*. Retrieved from https://app.lms.unimelb.edu.au/bbcswebdav/pid-6923739-dt-content-rid-45965483_2/courses/COMP30027_2019_SM1/2018-content/lectures/week09b.pdf.

Nicholson, J., Baldwin, T. & Verspoor, K. (2019). *COMP30027 Machine Learning Classifier Combination*. Retrieved from https://app.lms.unimelb.edu.au/bbcswebdav/pid-6923739-dt-content-rid-59980530_2/courses/COMP30027_2019_SM1/lectures/week07a.pdf.

Pal, M., & Mather, P. (2003, August 30). *An assessment of the effectiveness of decision tree methods for land cover classification*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0034425703001329.

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Saxena, S. (2018, May 12). Precision vs Recall. Retrieved from https://towardsdatascience.com/precision-vs-recall-386cf9f89488.

Singh, A. (2018, June 8). A Comprehensive Guide to Ensemble Learning (with Python codes). Retrieved from https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

Statnikov, A., Wang, L., & Aliferis, C. (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-319.

Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In Icml (Vol. 97, No. 412-420, p. 35).