**Objective**

The primary task in tax operation department is classifying the withholding tax incurred in each transaction payment. Normally, the process is to identify the service or work in those transaction based on the invoice description, purchase order and other supporting document. Then from those knowledge we identify whether it’s an object to tax liability and then determine the rate and the article of those tax liability.

This paper attempts to explore the possibility to derive the decision tree and decision rule in those classification process in hope it will open the opportunity to partially automate those process. The approach to derive the decision tree will be done using statistical learning (machine learning).

**Approach**

Statistical learning is subfield in statistic that the application objective is to extract information from data. Mainly there is two major category, supervised learning and unsupervised learning. Supervised learning is used when we have data feature or variable and its classification. Supervised learning is used to train the model using training data, which is historical data and use those model to predict the classification of future data which we haven’t yet know the classification. In summary the supervised learning is used to predict the future data based on previous data. The other type of statistical learning is the unsupervised learning is used when the data don’t have particular category or group and the learning process is aimed to group those data together based on particular characteristic in hope to gain meaningful information.

For this paper we will use supervised learning. The feature in those data is the description of transaction recorded in Oracle and the classification is the type of service and tax article applied to those transaction. In supervised learning there are many model that can be used to predict the the classification of future data but only decision tree and decision rule has the human readable reasoning process. In other model such as logistic regression, neural network, and SVM(support vector machine) the reasoning in the model is in form of coefficient in mathematical equation. So it hold no value in our purpose.

**Process**

In this paper the data we used is the one year historical data of tax payment. Ranging from Septemer 2015 to August 2016. The reasoning is because there is a major change in tax regulation at July and August where Telkomsel was appointed as WAPU (wajib pungut) and also appointed to withheld PPh 22 for the payment of goods which previously wasn’t an object of those tax liability. So it’s wise to use the data after those implementation so it will accurately reflect the current and possible future transaction data.

From those data we will use 2 primary feature, vendor name and invoice description. The classification group is the appended combination of tax article, tariff and type of service. For example if the tax article is PPh 23, its tariff is 2% and the type of service is “sewa dan penghasilan lain sehubungan dengan harta”, then the appended classification will be “PPh 23 2% sewa dan penghasilan lain sehubungan dengan harta”.

Before goes deeper into the process, first we need to evaluate what data we have in hand and the possible processing method suitable for those data. Mainly the processed data is in form of text and in human readable language, so we will use NLP(Natural Language Processing). NLP is a method to process the human readable text and communication form, such as email, written text message and audio language to the format processable by the computer for the statistical learning process. The result is a binary matrix that contain information whether those invoice contain particular word or not.

After the data is ready to be processed the algorithm will learn based on those data and evaluate which word affect most to the classification. To simply put the algorithm is split the data into two group based on whether it contain particular word or not and calculate where the entropy drop the most. Entropy here mean the variance (rate of error) in the data for those particular group. Word that drop the entropy most mean that those particular word is significant for the decision process. The process executed recursively until the variance reach bottom limit where there is no more room for improvement and we get all the keyword affected the classification process.

The result will be in the form of list of word that affect the decision process and the rate of error of those model. In complex transaction we realize relying the classification modelling only based on invoice description won’t be enough and the result won’t be optimum.