Quăng vô sau

## VietTreeBank[14]

In machine learning field, Treebanks are usually used as a tool to do most tag in the pre-process data step, such as training syntactic parsers, part-of-speech taggers, and word segmenters. For Vietnamese spicificly, VietTreebanks choose to use constituency representation of syntactic structures. There are three level of notation for VietTreebanks which include word segmentation, POS tagging, and syntactic labeling.

[These systems then can be used for applications such as information extraction, machine translation, question answering, and text summarization.] Since Vietnamese word order is quite fixed, VietTreebanks choose to use constituency representation of syntactic structures. Using annotation scheme proposed by Marcus et al. (1993). For Vietnamese, there are three annotation levels including word segmentation, POS tagging, and syntactic labeling. The goal is to build a corpus of 10,000 syntactically-annotated sentences (trees) and an additional POS tagged data set of 10,000 sentences. Raw texts are collected from a newspaper source, the Youth online daily newspaper (Tuoi tre online), with a number of topics including social and politics.VietTreeBank completed about 9,500 trees and 10,000 POS tagged sentences. For automatic labeling tools, they use CRFs for POS tagging or LPCFGs for syntactic parsing. In addition to treebank, other textprocessing resources and tools include: Vietnamese machine readable dictionary, EnglishVietnamese parallel corpus, word segmenter, POS tagger, chunker, and parser. Treebank and tools are closely related. Tools are trained using treebank data, and then they can be used in treebank construction.

* + 1. Word Segmentation in VietTreeBank

Dictionary editors may want to extract phrases from text which need to be explained in meaning. For this application, syntactic parsers can be used as tool for editors. Parsers can extract candidates for phrase/word entry. The following word types are considered in word segmentation phase: single words, compound words, repeated words, idioms, proper Vietnamese Language and Speech Processing names, date/time, number expressions, foreign words, abbreviations. Word segmentation ambiguity is the major problem annotators have to deal with. Suppose that three words “nhà cửa”, “sắc đẹp”, and “hiệu sách” are being considered. Annotators need to identify these combinations as words in: a. Nhà cửa bề bộn quá b. Cô ấy giữ gìn sắc đẹp. c. Ngoài hiệu sách có bán cuốn này And not words in: a. Ở nhà cửa ngõ chẳng đóng gì cả. b. Bức này màu sắc đẹp hơn. c. Ngoài cửa hiệu sách báo bày la liệt. They consider dictionary words as candidate for word segmentation and make decision using context.

* + 1. POS Tagging and Syntactic Annotation Guidelines

1. POS Tag Set

They choose first two criteria, combination ability and syntactic function, for POS tag set design. Therefore their POS tag set will not contain morphological information (number, aspect, tense, etc.), sub-categorization information (transitive/intransitive verbs, verbs followed by clauses, etc.), and semantic information.

1. Syntactic Tag Set

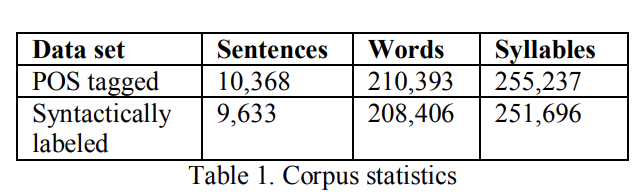
VietTreeBank tag set contains three tag types: constituency tags, functional tags, and null-element tags. They use the tag H to label phrase head. If a phrase has more than one head, connected by coordination conjunctions or commas, then all heads are labeled with H tag. Other treebanks often does not use head tag. Therefore researchers on syntactic parsing (Collins, 1999) used heuristic rules to determine CFG rules’ head. Machine learning methods also can be used (Chiang and Bikel, 2002). Null elements are often used for adjective clauses, ellipsis, passive voice, and topic.

1. Sentence and Phrase Analysis Techniques

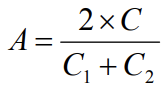
Annotation of real text requires various techniques to be applied. Ambiguity may occur in many steps of analysis such as determining phrase’s head, discriminating between possible complements, discriminating between adjuncts and other sentence elements, etc. Sentence analysis techniques include deletion, substitution, insertion, transformation, questioning. These techniques exploit contextual information, word combination, word order, and functional words to disambiguation between possible structures.

1. Annotation Process and Agreement

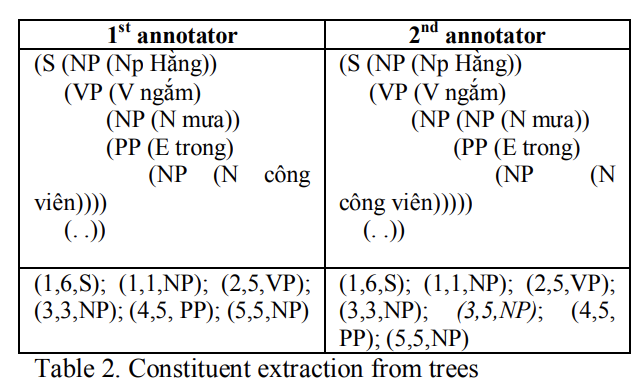
There are three annotation levels: word segmentation, POS tagging, and syntactic labeling. The word segmentation tool was used for the first annotation level (word segmentation) immediately. As to the other annotation levels (POS tagging and syntactic parsing), first several thousand sentences were labeled manually. After that a POS tagger and a parser are trained bimonthly, then the annotation task becomes semiautomatic. According to the annotation process, each sentence is annotated and revised by at least two annotators. The first annotator labels raw sentences or revises automatically-analyzed sentences. Then the second annotator revises the output of the first annotator. Table 2 shows a number of important corpus statistics such as sentence count, word count, and syllable count for two data sets. The average sentence length is about 21.6 words.



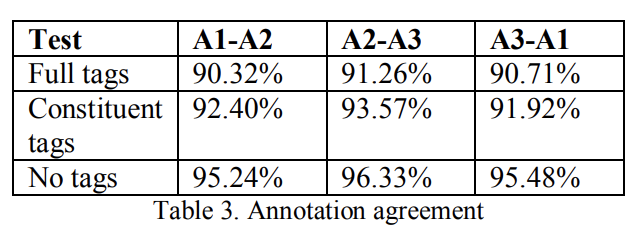
Annotation agreement measures how similar two texts annotated independently by different annotators are. Since this problem is similar to parsing evaluation, they use parseval measure. First, syntactic constituents in the form (i, j, label) are extracted from syntactic trees. Then tree comparison problem is transformed into constituent comparison. They compute three kinds of measurement: constituent and function similarity, constituent similarity, and bracket similarity. By using this method, they can evaluate both overall agreement and constituency agreement. Annotation agreement A between two annotators can be computed as follows:



where C1 is the number of constituents in the first annotator’s data set, C2 is the number of constituents in the second annotator’s data set, and C is the number of identical constituents. Table 3 shows an example of constituent extraction from trees. From Table 3, compute: C1=6; C2=7; C=6; A=12/13=0.92 .



They carried out an experiment involving 3 annotators. They annotated 100 sentences and the result is shown in Table 3

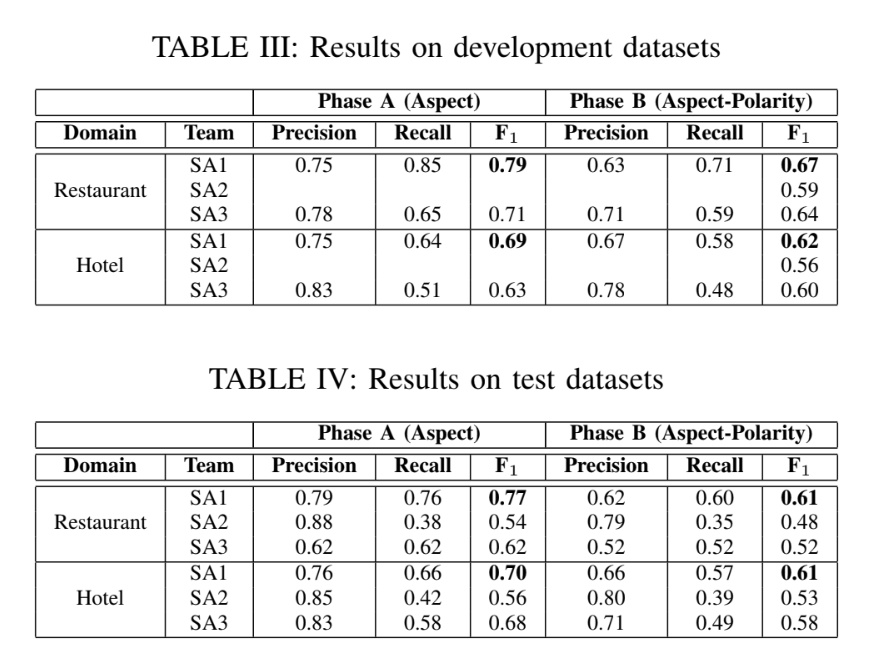


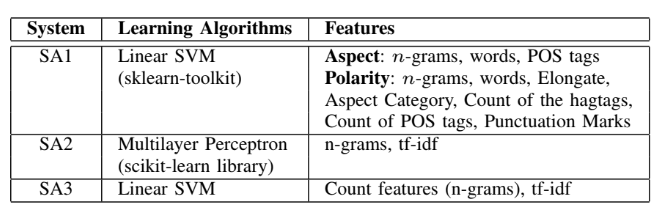
### **Vietnamese language and speech processing (vlsp) 2018**

With VLSP 2018, the goal is to identify a set of {aspect, polarity} tuples that summarize the opinions expressed in the review given a customer review about a target entity. Aspect is a pair of entity–attribute, while polarity can be “positive”, “negative” or “neutral” [11].

1. ….

The task considers reviews in two domains: Restaurant and Hotel. The subtask 2 of VLSP 2018 is focus on (Aspect-Polarity): The participants are required to identify both aspects and sentiment polarities. The dataset is taken from review site of hotel and restaurant. Using mostly SVM (more detail in the table) to build the model and predict. With the score in the table below:





While SA2 and SA3 considered the task as a multi-class classification problem (each label is a pair of aspect-polarity) and built only one classifier to solve the task, SA1 treated the task as multiple binary classification problems and built a single binary classifier for each aspect. To identify polarities of reviews, SA1 modeled the problem as a classification with three classes, i.e. positive, negative, and neutral.

Table II summarizes learning algorithms and features used in participating systems. While SA1 and SA3 used SVM with linear kernel, SA2 exploited multilayer perceptron algorithm. SA2 and SA3 built only one multi-class classifier with basic features, including n-grams and tf-idf scores. SA1 used more sophisticated features, such as elongate features, hagtags, punctuation marks. SA1 also conducted some preprocessing steps before training classification models.

1. Results

Tables III and IV summarize results of participating systems on development and test datasets, respectively. For both domains, SA1 achieved the best F1 scores on both development and test datasets. The results showed the effectiveness of sophisticated features used in SA1. Using linear SVM, SA1 and SA3 outperformed SA2 with multilayer perceptron significantly.

### **UIT-VSFC: Vietnamese Students’ Feedback Corpus for Sentiment Analysis**

The UIT-VSFC corpus is annotated for two different tasks: sentiment-based and topic based classifications. Ensuring the quality of corpus is one of the most important issues in the annotation process of the corpus. Therefore, they build a clear and complete annotation guidelines to control the annotation consistency and accuracy of the corpus. The annotation guidelines are designed for two separate tasks: sentiment-based and topic-based. The guidelines are not only useful for annotators but also important documents for researchers concerning sentiment analysis and education. [12]

The two key contributions are:

1) Obtain Vietnamese students’ feedback corpus (UIT-VSFC) for sentiment analysis and education research. As a result, they achieved over 16,000 sentiment-labeled and topic labeled sentences.

2) Analyzed the characteristics of Vietnamese students’ feedback and pointed out the challenges of annotating the corpus, which is crucial to build clear and complete annotation guidelines for annotators to control the quality of the corpus and support to build classification models on the corpus.

* + 1. Description

1. Task Definition

There are two subtasks: sentiment-based and topic-based.

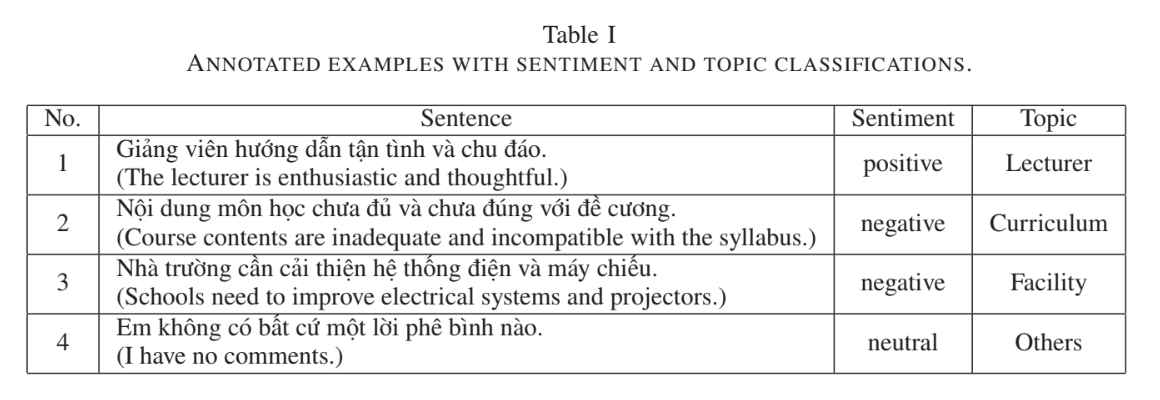
• Sentiment-based task:

Given a Vietnamese student’s feedback sentence, determine whether it expresses a positive, negative, or neutral/objective sentiment. For example, annotated sentences are shown in the Table I and the sentiment polarities are shown in the third column of the table respectively.

Each sentence is labeled one of three sentiment polarities including positive, negative, and neutral. Annotation guidelines are briefly described as follows.

* + Positive polarity: Sentences in which students use to express their satisfaction and compliments about elements of teaching activities such as lecturers, subjects, assignments, etc
  + Negative polarity: Sentences express students’ dissatisfaction, requests, and complaints related to Lecturer, Curriculum, Facility, and so on.
  + Neutral polarity: Sentences are incomplete or unclear in meaning or do not contain opinions, for instance, the sentence "giờ\_giấc và cách giảng dạy" (means: time and teaching methods) is a neutral sentence which is actually a noun phrase and does not contain any words expressing emotions. Or, sentences are complete but do not express any students’ feelings.

*Topic-based task*: Given a Vietnamese students’ feedback sentence, decide whether it expresses information related to Lecturer, Curriculum, Facility, or Others. The labeled examples with related topics which are shown in the fourth column of Table I respectively.



In this task, the author divided the topics which students are most interested in into four major classifications consisting of Lecturer, Curriculum, Facility, and Others:

*Lecturer:* Sentences express emotions about teaching methods, attitudes, knowledge, and so on. Detecting these feedback related to lecturers help to improve the quality of lecturers’ teaching activities. As an example, the given sentence "Thầy có phương pháp giảng dạy rất hay." (means: "You have a great teaching method.") portrays the compliment about a lecturer, therefore, this sentence is assigned to the Lecturer topic.

*Curriculum:* Sentences express the sentiment about matters related to subjects, curriculum, assignments, grades, session time, knowledge, labs, etc. For example, the given sentence "Môn\_học này rất cần\_thiết và bổ\_ích." (means:"This subject is very necessary and beneficial.") is assigned to the Curriculum category.

*Facility*: There are students’ feedback related to facilities such as computers, electric lights, projectors, fans, etc. For example, the given sentence "Máy tính phòng lab chạy quá chậm." (means: "The lab computers run slowly.") is labeled with Facility.

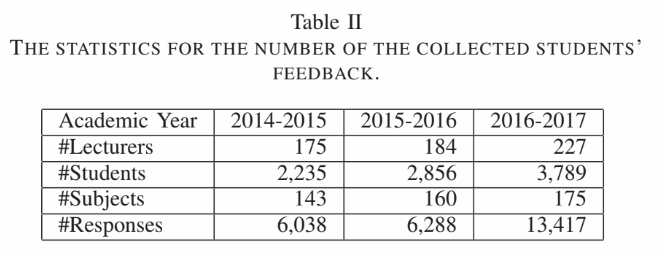
*Others:* Sentences, which are unclear or belong to other than the above topics, do not express any certain topics, so it belongs to the others type. For instance, the sentence "Em không hài\_lòng." (means: "I’m not satisfied.").

There are sentences which express both negative and positive emotions which are difficult to judge. They are often complex sentences with conjunctions such as nhưng (but), tuy nhiên (however), etc. In this cases, the annotators are asked to chose the stronger polarity based on their understanding. Example sentence "Thầy dạy dễ hiểu nhưng nhiều lúc hơi nhanh." (means: "The teacher’s taught was easy-to-understand but sometimes fast.") is assigned to the negative polarity, although it contains the positive clause "The teacher’s taught was easy-to-understand."

* + 1. Corpus building process

1. Collecting Data

Students’ feedback in the project were from a university in Vietnam, which were collected through student surveys by the university at the end of each semester from 2013 and 2016 with more than 16,000 students’ sentences. Table II shows the detailed information of the collected data.



1. Pre-processing

The normalization is done using NLP tools effectively such as POS taggers and parsers in the process of building models. Firstly, they segment student’s feedback into sentences. Then replace abbreviations with complete words and correct the misspelling. The abbreviations and spelling mistakes are stored in a dictionary for media data mining. As a result, the project obtain over 16,000 normalized sentences in this phase.

1. Annotation process

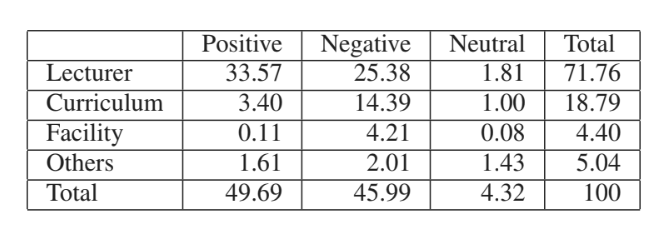
The author build the basic guidelines as completely as possible before starting the annotation process. They also update the annotation guidelines during the annotation process, which is necessary because there are many complicated cases. When finding a difficult case, they will update the guidelines. After updating the guidelines, they re-train the annotators with the revised guidelines. Eventually, they evaluate the corpus based on inter-annotator agreements and classification experiments.

D. The resulted corpus

They obtained over 16,000 sentiment-labeled and topic labeled sentences. Table III shows the distribution of sentiment and topic classes of the corpus.

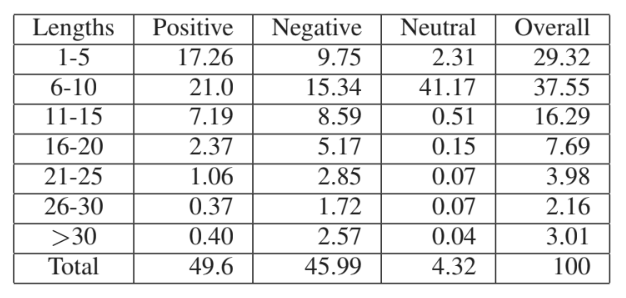
Table III

DISTRIBUTION OF THE CORPUS ON SENTIMENTS AND TOPICS (%).



The distribution of sentiment-based sentences according to the length of the sentence was presented in Table IV.

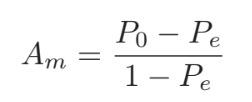
Table IV DISTRIBUTION OF SENTIMENT-BASED SENTENCES ACCORDING TO THE LENGTH OF THE SENTENCE (%).



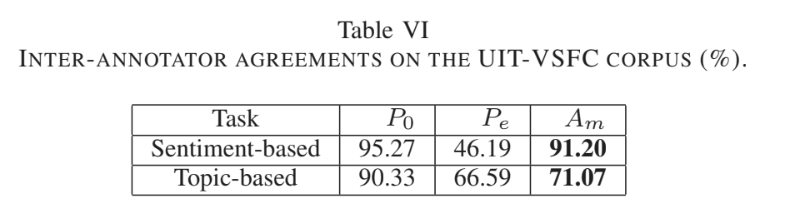
* + 1. Corpus evaluation

1. Inter-Annotator Agreements

In order to calculate inter-annotator agreements, they follow the agreement measure Am [20] which is defined as the proportion of agreement after the chance agreement is removed from consideration, given by the following formula.



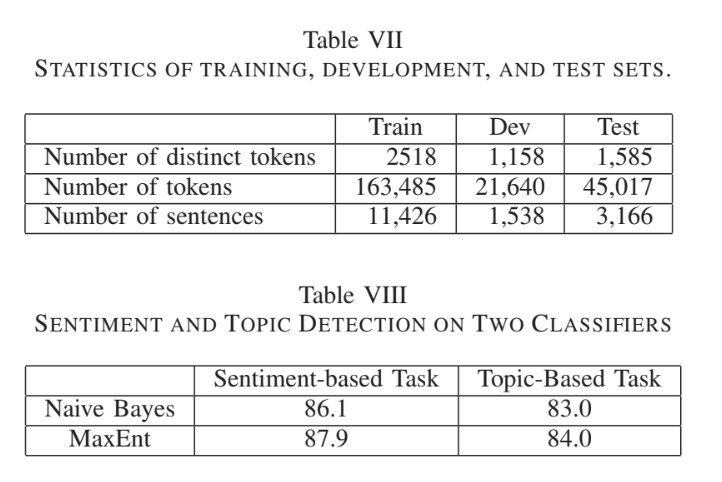
Where, the observed agreement (P0) is the proportion of sentences which both of the annotators agreed on the classes pairs and the chance agreement (Pe) is the proportion of items for which agreement is expected by chance when the sentences are seen randomly. Table VI shows the agreement measure of the corpus on three annotators for two tasks. With regard to the sentiments, the Am agreement and the observed agreement are very high with 91.20% and more than 95% respectively. Besides, the Am agreement of topics is 71.07%, whereas the topic observed agreement is higher than that with 90.33%. As a result, the agreement of sentiment is much higher than that of topics, over 20%. This implies that the annotators identify the sentiment polarities easier than the topics.



1. Experimental Results

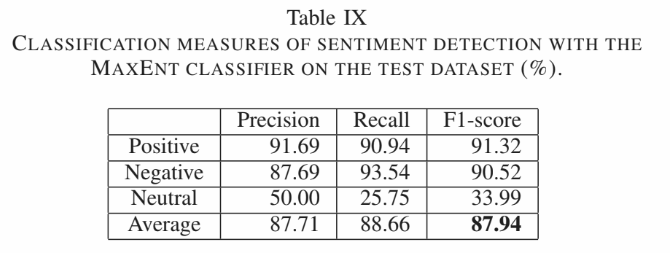
1) Setting: They implement two classifiers as Naive Bayes and Maximum Entropy on the corpus. In addition, they also use the five-fold cross-validation scheme. The result are calculate using four evaluation metrics such as accuracy, precision, recall and F1-score. To create the training, development, and test datasets, the original corpus was split randomly into three parts which cover approximately 70.0%, 10.0%, and 20.0% respectively.

Table VII shows statistics of these datasets in number of tokens and number of sentences. They used two available tools for classification experiments, the Datumbox framework [18] and the Stanford Classifier [17] for Naive Bayes and Maximum Entropy respectively.



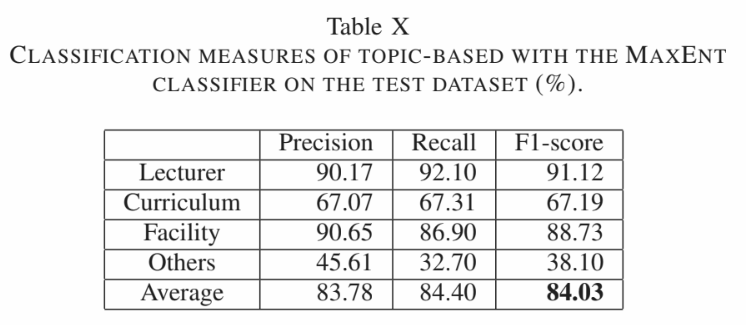
2) Sentiment Classification Results:

The performance of MaxEnt classification on sentiment-labeled data is better than that of Naive Bayes, as can be seen in Table VIII. The precision, recall and F1-score of the sentiment MaxEnt model on the test dataset are shown in Table IX. The results on both positive and negative are high with F1-score, 91.32% on the positive examples and 90.52% on the negative examples, whereas the performance on the neutral examples is low with the only 33.99%. This is because the corpus is unbalanced, for instance, the neutral data occupy 4.32% of the corpus.



3) Topic Classification Results:

The performance of Max-Ent classification on topic-labeled data is better than that of Naive Bayes, as can be seen in Table VIII. Table X shows the precision, recall and F1-score of the topic MaxEnt model. The result of the Lecturer class is high with the F1-score of 91.12%, whereas the figure for the Curriculum class are low with 67.19% of the F1-score. The overall measurement of topics is over 84%, lower than that of sentiments. In general, the quality of the classifications still needs further improvements.



## Introduction to Machine learning

The basic explanation of how a machine can learn is to have the results of the program developed after taking in Experience in a particular Task to meet the criteria set as Performance measure [20].

1. Tasks

The tasks in Machine Learning usually describe how Machine Learning processes the data point. In terms of image classification, the data point is one pixel in the image. Regarding customer classification, the data point is the customer and in the spam email classification, it is the email. The data point includes different features and each feature is described by number. Therefore, the data point is described by the vector x ∈ Rd, in which d: the number of features in the data point [21].

1. Performance measure

Performance measurement is needed to check the competence of the algorithms. When the Machine Learning algorithms are implemented, the dataset is separated into two parts: the training and the test set. The training set is used to find out the parameters, while the test set is used to measure the competence of the model.

In current models, the difference between the two sets is slightly ambiguous. According to Mitchell (1997), the parameters of the Machine Learning algorithms are always updated based on new data points, which is why those algorithms are called Online Learning. In the beginning, new data points are not used by the models, instead of using these data points to improve the parameters afterwards [22].

1. Experience

Training the Machine Learning model on the subset of the dataset can be considered to be the experience of a model on the training set. Thus, the quality of the training set affects substantially on the parameters of the model. Based on the properties of the dataset, Machine Learning algorithms can be classified into two groups: supervised learning and unsupervised learning [23].

First of all, Supervised Learning is the act of algorithms predicting the labels of the data points based on the data points training set [24]. Supervised Learning solves problems when the dataset has the input

and output , in which xiy i is the vector.

Based on the dataset, the function is built to map the x**i** to y**i** : ,

Secondly, in supervised learning, the output is unknown. The model is only built based on the input , which is a set of features of the data points. Based on the structure of the dataset, the unsupervised learning will build the model, for example, clustering or dimensionality reduction [25].

In the probabilistic view, unsupervised learning experiences the data points x to learn the probability distribution p(x). While the supervised learning experiences the input x and output y to optimize the conditional probability p(y|x)

Additionally, there are semi-supervised learning and reinforcement learning. In the dataset, the model based on to build includes the labelled and unlabelled data points. This type of learning occurs very frequently on the Internet when the small part of image or text is labelled, while the other large part retrieved from the Internet is unlabeled [26]. Many problems of Machine Learning today fall into this kind of learning because collecting the labelled dataset is a time-consuming and costly process. In addition, some datasets are too hard to classify such as medical images.

Meanwhile, Reinforcement Learning does not need the dataset to train. It experiences explicitly the environment, continuously receives feedback from the environment to improve, with the automatic driving car is the most popular application. Currently, reinforcement learning is applied to games so that the computer can simulate the environment [27].

1. Cost function and parameters

Model parameters are very important in each Machine Learning model. The purpose of the training process is to determine the most appropriate model parameters. Quantitatively, the best model parameters will make the best result in performance measurement. In the classification problem, the better the result in performance measurement is, the fewer misclassified points there are. Regarding the regression problem, the better the result gets, the less deviation between the predicted and target values there is.

The correlation between performance measure and model parameters is described by the cost function. The less the function costs, the better the model parameters are. Therefore, the training process in machine learning can be called the optimization process.

Suppose Θ: model parameters, L(θ): cost function of the model, then, Machine Learning models can be written by: Θ = arg Θ minL (Θ). Machine Learning model determines the Θ such that L(θ) is minimum [28]

1. Natural Language Processing (NLP)

Natural Language Processing, or NLP hereafter, is an automatical way to understand and process the information from natural language by human such as text or speech by software [29]. There are substantial problems which are solved by Machine Learning. The most common problems Machine Learning can solve are classification, regression, clustering, and completion.

Firstly, in the classification of the problem, it is required that machine learning must be able to predict the labels of the data point that are usually numbered from 1 to C. To solve the problem, the function: should be figured out. The difficulty of handwritten digit recognition MNIST is the most basic and popular classification.

In some current state-of-the-art models, the output of the classification problem is the vector y ∈ Rc in which yc is the probability of data point x has the label of c. Therefore, the label has the largest probability is the label for data point x.

Applying the probability in Machine Learning is very vital because it demonstrates the confidence of the model. In particular, if the largest probability is considerable compared to the second largest probability, it can be said that the model is high confidence in the classification of data point x. On the contrary, if the difference between the largest and the second largest probability is small, it’s likely to misclassify [30].

Secondly, in the regression problem, the output of the model is the real value instead of the label. The function:  . An example of that is to estimate the price of the house which has the area of x (m2), y bedrooms and the distance from downtown is z (km).

Especially in the advanced regression problems, the output of the model is multiple real values, so the function in this case is : . Take a single image with super-resolution, the model in this problem make the higher resolution image based on the lower resolution image.

Thirdly, clustering will classify the unlabelled datasets X into subsets based on the correlation among the data points in one subset, for instance: clustering customers based on their buying behaviors. This application is very popular and assists a lot in recommender systems.

Finally, completion is the problem of filling the missing data points. In real datasets, collecting all information is almost infeasible, since collecting all ratings of all users on all items in the recommender system is too difficult because most of the users do not explicitly rate the items on the website. Therefore, the completion will fill the missing data points based on the correlations among the data points.