University of Alberta

Automated Essay Scoring Framework for a Multilingual Medical Licensing Examination

by

Syed Muhammad Fahad Latifi

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation and Cognition

Department of Educational Psychology

Spring 2014

Edmonton, Alberta

Abstract

Automated essay scoring (AES) is a technology that efficiently and economically score written responses by emulating intelligence of human scorer. Present study had employed open-source Natural Language Processing technologies for developing AES framework, to score multilingual medical licensing examination. English, French, and translated-French responses of constructed-response items were scored automatically, and the strength of multilingual automated scoring framework were evaluated in relation to human scoring. Machine-translation was also contextualized for raising AES performance, when restricted sample size counters the performance of AES software. Specific feature extraction and model building strategies resulted in high concordance between AES and human scoring, with average maximum human-machine accuracy of 95.7%, which was in almost perfect agreement with human markers. Results also revealed that the machine-translator had raised predictive consistency but negatively influenced the predictive accuracy. Implications of results for practice, as well as directions for future research are also presented.

Acknowledgements

I like to thank all great minds with whom I worked during this research. Specially, I want to thank my research supervisor and academic mentor Dr. Mark J. Gierl, for his advice, his encouragement, and mentorship. Thank you Dr. Gierl for everything throughout this project, and throughout my Master of Education training, you are an inspirational leader. I would also like to thank the members of my thesis committee, Drs. Lai, and Cormier, for their expertise and advice.

I extend my gratitude to the Medical Council of Canada, Ottawa, for providing me the opportunity to carry out this research. Specifically, to Mr. André-Philippe Boulais for providing me the information to establish this research.

Above all, I like to thank my mother and father(late) for everything they did for me, and being the source of support and encouragement in phases of my life, I cannot find words for thanking them.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Writing is one of the most powerful methods for assessing 21st century skills such as critical thinking, problem solving, communication, creativity, and innovation (Rich, Schneider, & D'Brot, 2013). Considerable resources are now being channeled towards measuring writing ability as evidence of academic skill acquisition. As a result, there is an increased demand to develop efficient assessment systems that can measure the higher-order thinking and writing skills of students (Attali, Lewis, & Steier, 2013; Shermis & Hamner, 2013). However, such assessments often require long-written responses, and consequently, they are difficult to score in an efficient, economical, and objective manner.

One possible solution to address these problems is the technology of *automated essay scoring (AES)*. AES uses computer models (or scoring models) to score student-produced writings (Brew & Leacocle, 2013; Shermis & Burstein, 2003). An AES program is a computer software that builds the scoring models from pre-scored essays, using artificial intelligence and machine learning approaches, and then uses these models to grade new sets of essays automatically (Schultz, 2013). AES offers many exciting benefits for writing assessments, such as improving the quality of scoring, reducing time for score reporting, minimizing cost and coordination efforts for human raters, and the possibility of providing immediate feedback to students on their writing performance (Myers, 2003; Weigle, 2013; Williamson, 2013).

Background to Problem

The early idea of scoring students' writing using computers was advanced by Ellis Page in 1966 (see Page, 1966) with his AES program called *Project Essay Grade*. However, it was not until early 1990s when the advancement in automated scoring technology was resurrected (Keith, 2003), followed by the rapid revolution in technological innovation when assessment professional began to embrace and integrate these technologies into student assessments. This rapid growth in learning technologies has also given rise to commercialization and technology proprietorship. And the development of AES is not an exception. The information technology involved in automated scoring programs largely exists in the proprietary domain (Elliot & Klobucar, 2013). Researchers and licensing authorities who wish to study and automate the scoring of constructed-response items can incur significant start-up costs, only to establish a proof-of-concept.

Fortunately, researcher at the *Language Technologies Institute* of Carnegie Mellon University recently released a free and open-source machine learning environment called LightSIDE (*Light S*ummarization *I*ntegrated *D*evelopment *E*nvironment). This program has a user-friendly interface and it incorporates numerous options to develop and evaluate machine learning models. These models can be used for a variety of purposes including AES, thereby providing open-source access to otherwise proprietary technology. More excitingly, in a recent AES comparative study, LightSIDE performed equally well compared to other proprietary automated scoring programs developed by large testing companies (Shermis & Hamner, 2012; Shermis & Hamner, 2013). However, little is known about how well LightSIDE would perform if different machine learning algorithms (*MLAs*) were applied to score responses from different context. One example being, the written responses from the medical licensing context (e.g., questions which assess examinees' knowledge in medical clinical decision-making situations).

In order to test these machine learning algorithms, the structure of the writing passages also needs to be considered. For instance, Shermis and Hamner (2012) showed that different AES programs have different efficiencies for scoring essays with different text features (e.g., length of writing passage, number of words). While long essays can tap the maximum capacity of an AES program, they also introduce more ambiguities for researchers to define the underlying scoring process. Consequently, it is important to test the efficiency of machine learning algorithms by considering the properties of the input data for multilingual essay responses.

Purpose of Current Study

The research questions in this study were motivated to address three interrelated issues corresponds to the design and application of an automated scoring system within an open-source technology. That is, my study was designed to address the gap in the current academic literature describing the open-source AES framework available to researchers and licensing authorities who wish to study and automate the scoring of student-produced written responses. My study helps establish the proof-of-concept for an automated scoring process within an operational testing program. My study is also intended to contextualize the machine-translation process in an educational testing environment using automated scoring technology. The aim of the present study was to fill these research gaps and to convey the results that are of interest to a broad audience of researchers and practitioners. My study therefore addresses academia, assessment and licensing authorities, and researchers who wish to explore and advance the open-source automated essay scoring technology. Thus, the purpose of this research is threefold:

1) to develop and demonstrate the automated scoring framework using open-source Natural Language Processing (NLP) technologies for a multilingual medical licensing context;
2) to compare the strength of three multilingual score prediction engines with scores obtained from human raters; and
3) to contextualize machine-translation, in automated score prediction process, for raising the performance gain when restricted sample size counters the model building process.

To begin, related literatures are reviewed in Chapter 2. This review is then followed by the methods in Chapter 3 and the results in Chapter 4. Conclusions, discussion, limitations, and directions for future research are presented in Chapter 5.

Chapter 2: Literature Review

In this chapter, an overview of automated score prediction algorithms are presented. Three machine learning algorithms (MLAs) are reviewed, namely Naïve Bayes, Sequential Minimum Optimization, and J48-program. For a detailed technical review, corresponding references are also appended. Quantitative measures of automated scoring accuracy and consistency are then described.

Types of Prediction Algorithms

Automated score prediction algorithms fall in two broad types, supervised and unsupervised algorithms. The supervised algorithm is a machine learning algorithm that uses the pre-scored training samples (essays) to learn the approximated behaviour of human scoring process by looking at several examples of pre-scored essays. This step is called the model building/learning process. The built model could then be used for automated scoring of a separate or a new set of essays based on the likelihood suggested by the regression steps of the model (Yannakoudakis, Briscoe, & Medlock, 2011).

The unsupervised algorithm is a machine learning algorithm that does not require any pre-scored reference samples (essays) for building the learning model. This approach is called unsupervised because there is no need for human intervention and/or the requirement of supplying the pre-scored essay documents at any point in the process (Lee & Yang, 2009). For unsupervised algorithms, learning is based on the content of the individual essays and their divergence from the collection of essays, where the collection is considered as one large essay (De & Kopparapu, 2011). Empirically, the unsupervised learning algorithms are less accurate than the supervised learning algorithms. That is, the trade-off is the less accurate score prediction model for unsupervised learning verses the expensive task of acquiring the pre-scored training dataset for supervised learning (Xiong, Song, & deVersterre, 2012).

Most automated scoring systems are based on supervised learning algorithms (De & Kopparapu, 2011). Ironically, the machine learning methods used for automated scoring process are not described in any detail in the published work of AES systems (Yannakoudakis, Briscoe, & Medlock, 2011). For this study, three supervised learning algorithms were used and each algorithm had a distinct approach for building the score prediction models. The theoretical overview of each learning algorithm is presented next.

Overview of Naïve Bayes

The Naïve Bayes is a probabilistic classification method based on Bayes' theorem and the assumption of conditional independence (Mitchell, 1997). The main question Naïve Bayes classifier tries to answer is: given a subject has a certain set of features, what is the most likely class that this subject belongs to? For example, given a fruit is red, round, and about 7 centimeters in diameter, is this fruit more likely to be an apple or an orange? In order to solve this problem, the Naïve Bayes method estimates the conditional probabilities for both apple and orange, and chooses the class that has the highest conditional probability. To estimate these conditional probabilities, the Naïve Bayes classifier makes a strong conditional independence assumption: within each class (i.e., apple or orange), the features (i.e., roundness, color, and diameter) are independent of one another.

This means Naïve Bayes assumes that the roundness of an apple is unrelated to its redness or its diameter. In practice, this assumption is often violated. Therefore, the term “Naïve” is used to remind us of this assumption. That said, many empirical comparisons between Naïve Bayes and other more complex classification algorithms show that Naïve Bayes is simple, efficient, and provides comparable performance (Chen, Huang, Tian, & Qu, 2009; Kononenko, 1990; Pazzani, 1996). To put the above example into mathematics, Naïve Bayes classifier is expressed as:

$$argmax_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c), \qquad \ldots..(1)$$

where$f_1, \ldots, f_n$ represent the features and *n* indicates there are a total of *n* features; $p(C = c)$ represents the probability of a class *c* under class variable C (e.g., what is the probability that a random fruit is an apple?); $p(F_i = f_i | C = c)$ represents the conditional probability that given a case belong to class *c*, what is the probability of having the i*th* feature F_i equals to value f_i (e.g., given the fruit is an apple, what is the probability that its color is red?).

In practice, $p(C = c)$ and $p(F_i = f_i | C = c)$ are first estimated from a training dataset that contains examples that have already been classified. Then, the estimated $p(C = c)$ and $p(F_i = f_i | C = c)$ can be input into equation (1) to classify a new dataset.

Overview of Sequential Minimal Optimization

Unlike Naïve Bayes classification, Sequential Minimal Optimization (SMO) by itself is not a classification method. However, SMO can be considered as a part of a classification method called *Support Vector Machine* (SVM; Platt, 1998). While SVM and Naïve Bayes address similar classification problems, SVM uses a different approach. Instead of using a probabilistic approach, SVM uses a geometric or linear algebraic approach. Basically, SVM can be conceptually understood as representing subjects as points in space, which is mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible (Platt, 1998; Tong & Koller, 2002). This basic idea can be illustrated by Figure 1, which represents a simple classification problem: given the values of two features (i.e., F_1 and F_2) of a subject, what class (c1, c2) does this subject belongs to?

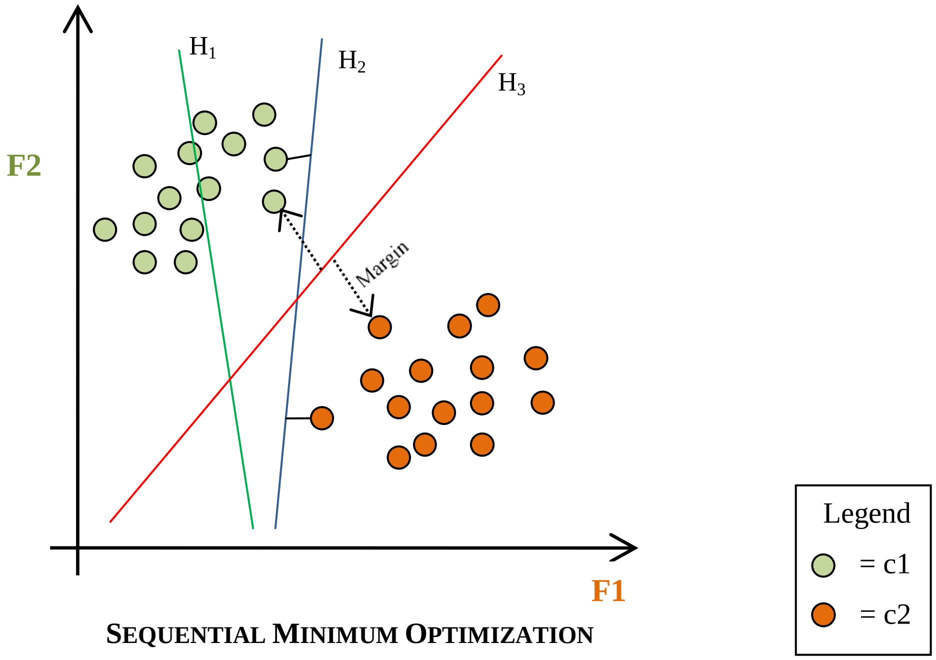


Figure 1. Conceptual representation of sequential minimum optimization.

As shown in Figure 1, the solution to this problem is to find a line (or hyperplane when there are more than two features) that best separate the two classes c1 and c2 by a maximum margin. The criterion for best separation is that the distances from the nearest point to the separating line must be a maximum. As shown in Figure 1, the line *H1* cannot separate c1 and c2; line H2 separates c1 and c2, but the distances from the nearest point to H2 is small; line H3 separates c1 and c2, and the distances from the nearest point to H3 is at its maximum.

Mathematically, this geometric problem can be generalized and represented by the following optimization problem:

$$Max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_jK\left(x_i,x_j\right)\alpha_i\alpha_j \ldots\ldots (2)$$

Subject to: $0 \leq \alpha_i \leq H,$ for $i = 1, 2, \ldots, n,$ and $\sum_{i=1}^{n} c_i\alpha_i = 0,$

where *n* is the number of subjects; x_i is a vector that contains all the feature values for subject *i*; c_i is the class for subject *i*; α_i is the Lagrange multipliers; H is an SVM hyper parameter; $K(x_i, x_j)$ is a symmetric *Mercer kernel function* (Souza, 2010; Tong & Koller, 2002).

The SMV-based algorithm, SMO, is an efficient algorithm to solve the optimization problem presented in Figure 1. Conceptually, SMO splits the classification problem into a series of the smallest possible sub-problems, and then solves these problems analytically (see Platt, 1998 for a detailed mathematical explanation).

Overview of J48 Program

The J48 program is an open-source java implementation of the C4.5 classification algorithm (Costa et al., 2013). The C4.5 algorithm addresses the classification problem using a slightly different approach compared to Naïve Bayes and SMO. It uses the information gain approach (Costa et al., 2013; Quinlan, 1993). The basic idea for this algorithm is to split the data into subsets based on a feature that offers the most information gain, and then split the subsets based on another feature that offer the most information gain at the subset level. This process is repeated until all the features are used. For the previous example of apple versus orange, the first step is to determine what feature (i.e., shape, color, or diameter) offers the most information gain for the classification of fruit. Conceptually, since apple and orange have similar diameter and shape, splitting the dataset based on these features will not lead to much information gain. But if the dataset is split based on color, we would expect most of the red colored fruits to be apples and most of the orange colored fruits to be oranges. Therefore, in this case, color is the feature that provides the most information gain for the classification.

Mathematically, this information gain can be calculated using equation 3:

$$Information\ gain =$$

$$-\sum_{c=1}^{k} P(c) log_2 P(c) - \sum_{f=1}^{l} (\left(\frac{|S_f|}{n}\right) * \left(-\sum_{c=1}^{k} P(c_f) log_2 P(c_f)\right), \quad \ldots\ldots(3)$$

where c is the class value; k is the total number of class; P(c) is the proportion of c in the whole sample; f is the value of a feature; l is the total number of values in a feature; $|S_f|$ is the number of elements in a subset which has a feature value of f; n is the total sample size; $P(c_f)$ is the proportion of c in a subset that has a feature value of f.

Equation 3 can be applied repeatedly to determine the next best feature that leads to the most information gain. After the next best feature is determined, such as shape in the fruit example, the subsets can be further split based on this feature. C4.5 also includes some specific algorithms to deal with continuous features and missing data (see Quinlan 1993 for detailed explanation of J48).

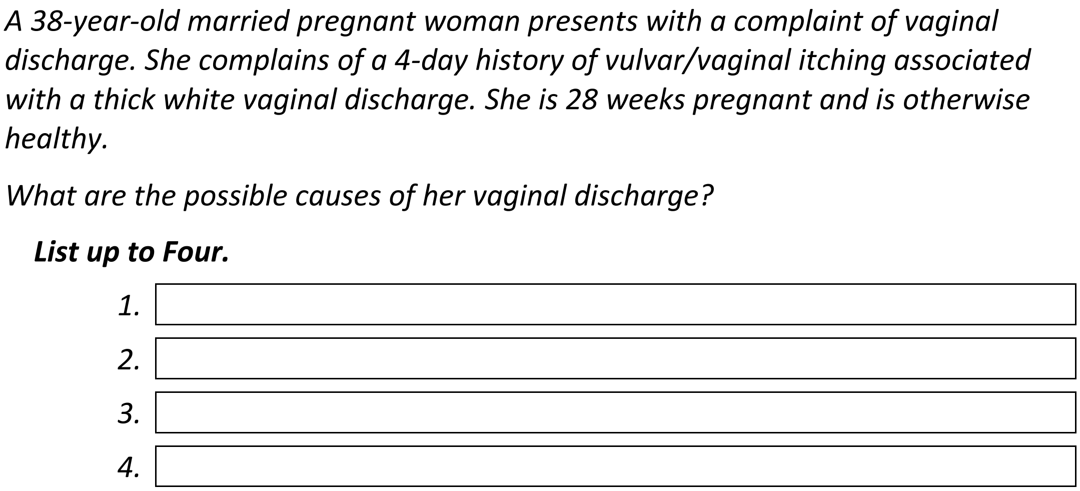
Model Performance Measures

The performance of machine learning algorithms in predicting human scores is evaluated using measures of accuracy and consistency. For accuracy, the exact-agreement percentages between human-produced and machine-predicted scores are computed. To measure the consistency between human and predicted scores, Cohen's (1960) un-weighted Kappa is computed. The interpretation of these performance measures is discussed in chapter 3. Methods of this study are presented next.

Chapter 3: Method

Dataset for Automated Scoring

Pre-scored responses from eight clinical decision-making (CDM) construct-response (CR) items were extracted from the examination system of a national medical licensing authority. The CDM items are a component of the licensing exam in which a clinical situation is presented to examinees and then they are expected to respond with a correct set of diagnoses, treatments, or prescriptions. The examinees could answer CDM items either in English or in French. The CDM component of the licensing exam carries two types of CDM-CR items. The first type ask an examinee to type-in the short answers for the given clinical situation while the second type ask an examinee to select as many of the given choices as appropriate. A sample of these CDM-CR questions are presented in Figure 2 and Figure 3, respectively. For the purpose of this research, eight CDM-CRs items that requires examinees to type-in the short answers were used.



A 38-year-old married pregnant woman presents with a complaint of vaginal discharge. She complains of a 4-day history of vulvar/vaginal itching associated with a thick white vaginal discharge. She is 28 weeks pregnant and is otherwise healthy.

What are the possible causes of her vaginal discharge?

List up to Four.

1. □
2. □
3. □
4. □

Figure 2. Sample CDM-CR item that requires to write-in short answers.

An anxious young mother brings her 14-month-old daughter to the Emergency Department. The child has had diarrhea for 3 days for which her mother gave her dimenhydrinate and apple juice. The child has not vomited. She appears ill and is crying with no tears in her mother's arms. She seems weak and lethargic. Her vital signs are:

Temperature	*36.5°C (axillary)*
Pulse	*160/minute*
Respirations	*70/minute*
Blood pressure	*95/35 mm Hg*

She weighs 13.5 kg. She has no tearing. Mucous membranes are dry to the touch. The capillary refill time is abnormal. There is diminished skin turgor with some tenting of the abdominal skin. The diaper is dry except for a small amount of watery green stool.

How should you manage this child?

Select as many as are appropriate.
(N.B. There are 12 options.)

1. □ *Abdominal radiograph*
2. □ *Apple juice and monitoring of intake*
3. □ *Arterial blood gases*
4. □ *Blood glucose*
5. □ *Blood urea*
6. □ *Complete blood count*
7. □ *Flat ginger ale and monitoring of intake*
8. □ *Glucose water and monitoring of intake*
9. □ *Intravenous infusion of 5% dextrose*
10. □ *Intravenous infusion of 5% dextrose in normal saline*
11. □ *Nasogastric rehydration solution*
12. □ *Nothing by mouth*

Figure 3. Sample CDM-CR item that requires examinees to select as many of the given choices as appropriate.

An overview of CDM-CR scoring. Typed CDM-CR responses were presented to human markers on a computer screen. The markers are appointed physicians who have expertise in the subject matter. They refer to well-established scoring rubric for assigning partial or full credit to the examinee`s CDM response. A scoring rubric is a criterion for assigning a score to the CDM-CR responses. There is a maximum of one mark per CDM question and an examinee`s response could also secure partial score. For instance, a CDM question whose scoring criterion includes four elements of a correct answer allows an examinee to receive partial marks, which in this case will be 0.25, 0.5, 0.75, or 1. For the purpose of this research, all eight CDM-CR items had a maximum score of 1 and minimum score of 0, with no partial credit between these score points.

Each CDM-CR response was scored individually by two human markers. As a result, each CDM response is scored twice. The scores were then compared for agreement and the final score is determined if there is exact agreement between two markers. In a case of disagreement between scores, the markers review the response in pairs to resolve the disparity in their score assignment. As a result of the in-pair review, the final score for discrepant cases were determined after exact agreement between markers.

Sample Size for CDM-CR items. For both language groups, English and French, pre-scored responses of eight CDM-CR items were extracted from the spring 2011, 2012, and 2013 exam administrations. Responses from the spring 2011 and 2012 exam administrations were used for feature extraction, training, and model validation processes. The validated models were then used for automatically scoring the responses from spring 2013. The anonymity of examinees was preserved by omitting names and demographic information. The sample sizes for training and automatic score prediction are presented in Table 1.

Table 1

Sample Size Summary for Training and Prediction Dataset in English and French

CDM-CR Question	***Training Sample (Dataset-I)***		***Prediction Sample (Dataset-II)***	
	English	French	English	French
1	1266	226	856	149
2	1259	237	1268	241
3	1255	238	859	143
4	1269	226	1256	246
5	1255	239	1256	246
6	1205	254	1272	241
7	827	144	848	156
8	1241	242	1260	228

In the subsequent sections, the spring 2011 and 2012 CDM-CR responses, which were used for feature extraction, training, and model validation, will be referred as *Dataset-I* while the spring 2013 CDM-CR responses, which were used for automated score prediction process, will be referred as *Dataset-II*.

Developing and Evaluating Classification Models

To develop the work-flow for automated essay scoring, the computer program LightSIDE (**Light S**ummarization **I**ntegrated **D**evelopment **E**nvironment) was used. LightSIDE is a open-source machine learning software environment written in Java. The performance of LightSIDE was tested and found to be remarkably high on essay scoring tasks, matching or exceeding human performance nearly universally (Mayfield & Rose, 2013; Shermis & Hamner 2013). LightSIDE offers many *MLA* implementations using machine learning software libraries from Weka.

Weka, an acronym of **W**aikato **E**nvironment for **K**nowledge **A**nalysis, is developed and updated by the researchers at University of Waikato, New Zealand (Hall et al. 2009). Weka is a comprehensive suite of software libraries which provides a general-purpose environment for text classification, regression, clustering, and feature selection which are collectively referred as data-mining (knowledge-discovery) and machine learning (Frank et al., 2005; Frank, Hall, Trigg, Holmes, & Witten, 2004; Witten, et al., 1999). Software libraries of Weka are also programmed using Java, and are freely available for import and extension into other software environments, such as LightSIDE. To develop and evaluate the automated score prediction models using LightSIDE, three MLAs (Naïve Bayes, SMO, and J48) were used. The design of this research is shown in Figure 4 and detailed steps are presented next.

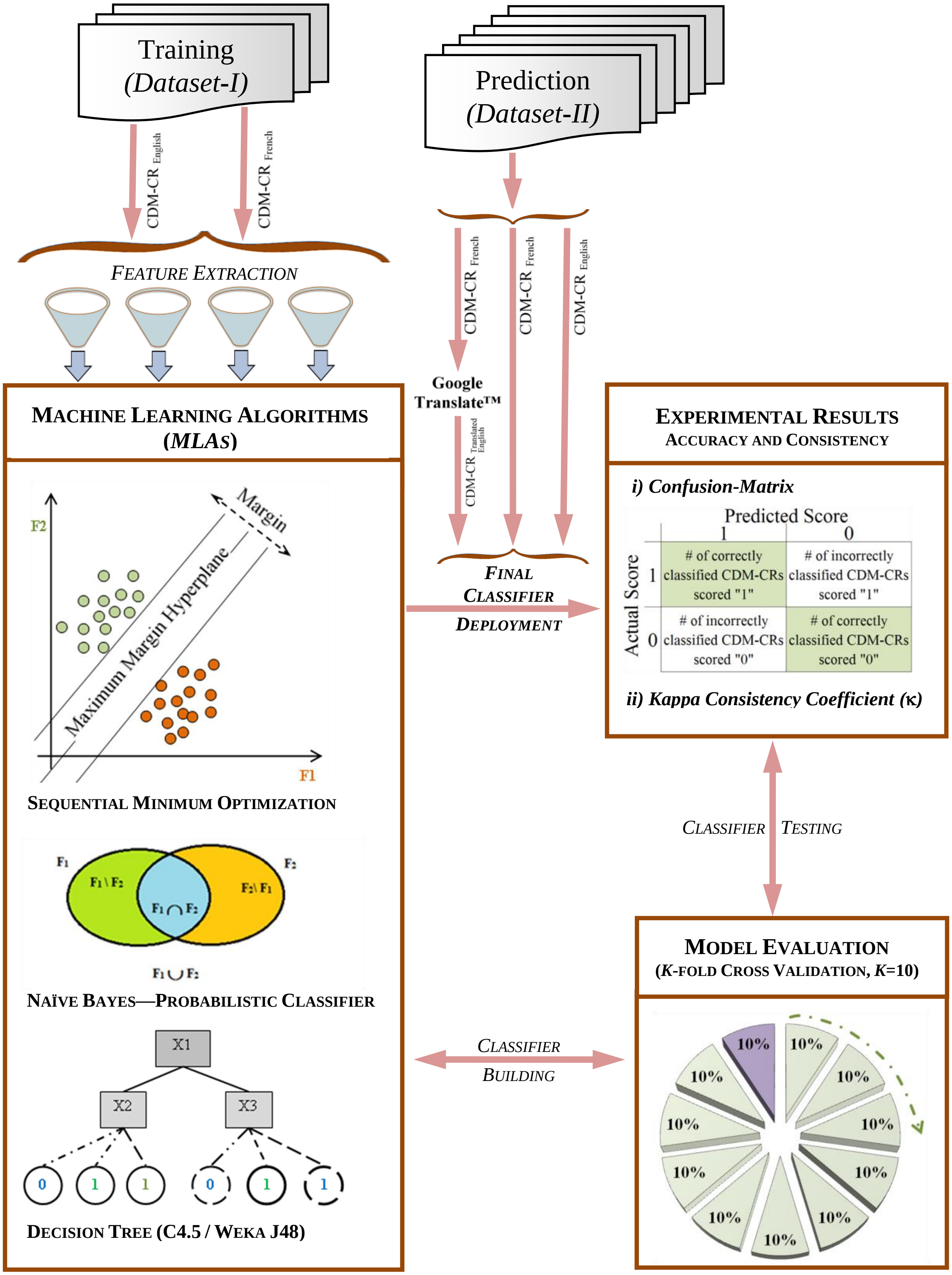

Training
(Dataset-I)
CDM-CR English
CDM-CR French
Prediction
(Dataset-II)
FEATURE EXTRACTION
CDM-CR French
CDM-CR French
CDM-CR English
Google
Translate™
CDM-CR Translated English
MACHINE LEARNING ALGORITHMS
(MLAs)
F2
Margin
Maximum Margin Hyperplane
F1
SEQUENTIAL MINIMUM OPTIMIZATION
F1
F2
F1 \ F2
F1 ∩ F2
F2 \ F1
F1 ∪ F2
NAÏVE BAYES—PROBABILISTIC CLASSIFIER
X1
X2
X3
0
1
1
0
1
1
DECISION TREE (C4.5 / WEKA J48)
FINAL
CLASSIFIER
DEPLOYMENT
EXPERIMENTAL RESULTS
ACCURACY AND CONSISTENCY
i) Confusion-Matrix
Predicted Score
1
0
Actual Score
1
# of correctly classified CDM-CRs scored "1"
# of incorrectly classified CDM-CRs scored "1"
0
# of incorrectly classified CDM-CRs scored "0"
# of correctly classified CDM-CRs scored "0"
ii) Kappa Consistency Coefficient (κ)
CLASSIFIER TESTING
MODEL EVALUATION
(K-FOLD CROSS VALIDATION, K=10)
CLASSIFIER
BUILDING
10%
10%
10%
10%
10%
10%
10%
10%
10%
10%

Figure 4. The research design of automated scoring framework for multilingual medical licensing examinations.

Steps for Creating Automated Scoring Models. Automated essay scoring is based on text classification (Rudner & Liang, 2002), which employs three basic steps for classifying texts (i.e., scoring essays) automatically. The first step is called the Feature Extraction process, which extract statistical relationships among elements-of-text (i.e. training dataset), thereby emulating the indirect relationship between elements-of-text and writing quality (Attali, 2013). The second step, often called Model Building and Self-Evaluation, employs the extracted text features to train, build, evaluate, and rebuild the scoring model. The third step, called Model deployment, which employs the final models from step two, on specific subject matter, and could be applied to a large pool of essays for automated text evaluation. Details for each of these steps are presented next.

Step One – Data Pre-processing and Feature Extraction

Data pre-processing. For each CDM-CR items, the responses were pre-processed before passing on to the feature extraction module. Two sub-steps are required. First, raw scores from markers must be annotated into labels. That is, each score point is replaced by text scoring labels. For instance, if a response is scored 1, then the annotated score label will be "one" (a text-label for number 1). This annotation is a requirement for the machine learning (classification) process. Second, the input dataset must also be cleaned from separators (e.g., commas ',' or pipelines '||' symbols). After annotating and removing separators, the CDM-CR responses were converted into a CSV file format for the feature extraction process using LightSIDE.

Statistical Feature Extraction. Across both linguistics groups, the item wise CDM-CR responses from *Dataset-I* were used to extract text features. Feature extraction is a process of objectively transcribing the input pool of texts (i.e., CDM-CR responses) into corresponding *n-gram*, *line length*, and *non-stop words*. *n*-gram is a sequence of terms with length *n*, the word of length one, two, and three were extracted and thus could be named as unigram, bigram, and trigram, respectively. Features beyond trigrams were not extracted as it could reduce the performance of the classification model (Fürnkranz, 1998). *Line length* feature creates a single numeric feature representing the counts of words in an essay instance. A *non-stop word* feature creates a single Boolean feature representing whether an essay instance carries any content-based words. *n*-gram were ignored if they did not qualify with the threshold of five occurrence in the training dataset. Each feature was extracted as binary feature; that is, the features are encoded as presence or absence of a particular word (or *n*-gram) among the training dataset. Further, the *stop-words* were removed from the CDM-CR responses. *Stop-words* are function words like “and” or “are” which do not contribute to the content of training responses (i.e., CDM-CR responses) because scoring a CDM-CR response is more about content than style (i.e., examinee`s writing style has no credit in the scoring rubric). As a result, removing *stop words* will improve the prediction accuracy of the automated scoring model (Dave, Lawrence, & Pennock, 2003; Mayfield, Adamson, & Rose, 2013; Pitler, Louis, & Nenkova, 2009). Item-wise extracted features, of English and French CDM-CR responses, were then used for the model building and evaluation process.

Step Two – Model Building and Self Evaluation

Model building. *Model Building* is an iterative process in which the extracted features were passed to the MLA so that the learning patterns and associations among the elements of texts could be analyzed. The objective of this iterative process is to develop a text classifier capable of intelligently classifying (i.e., scoring or grading) the written prompts. The classifier building process starts by judging the statistical features of text and building its own knowledge base or classification model from the training dataset (i.e., *Dataset-I*). While the classifier is learning from the text features, the performance of a classifier is measured in terms of prediction error (Rodriguez, Perez, & Lozano, 2010). In most real-world problems, the error cannot be calculated exactly and it must be estimated. Therefore, choosing an appropriate estimator for model prediction error that can validate the developing model requires some considerations.

Model self evaluation. The validity of classification model could be evaluated in two ways: validating using separate datasets or validating by splitting and re-using the training dataset, which is also called cross-validation. Acquiring separate datasets for the purpose of model evaluation is often difficult because large amounts of unique data must be available. In the current study, only a small amount of data were available. Therefore, cross-validation was employed. For splitting the data into training and testing dataset, one can choose to make a single split (e.g., half of the data for training and other half of the data for evaluation) or multiple (*K*) splits, which is commonly referred as *K*-fold cross-validation. Model evaluation by means of *K*-fold cross-validation is practical because it is computationally feasible (Arlot & Celisse, 2010) and economical because it does not require additional data. However, the computational cost is depended on the number of splits and the model statistical performance often does not improve if the number of splits exceed ten (Arlot & Celisse, 2010; Kohavi, 1995). For the purpose of this research, model evaluation by means of *K*-fold cross-validation (Geisser, 1975) was employed and training dataset were randomly split into *K* subsets for the model building and evaluation process.

K-fold is an attempt to estimate how accurately a scoring model would perform during the actual score prediction process. The model evaluation using *K*-fold cross-validation is achieved by randomly splitting training dataset, *Dataset-I*, into *K* mutually exclusive datasets of approximately equal size. The *MLA* is then trained and tested using *K* datasets (i.e., folds) in *K* iterations. In each iteration, the MLA is trained on all but one fold and tested on the remaining single fold. The overall accuracy of a trained MLA (i.e., model) is computed by averaging the *K* individual accuracy measures. For the purpose of this research, ten-fold (*K*=10) cross-validation is employed for model evaluation. Mathematically this could be represented as:

$$\Upsilon = \quad \frac{1}{K}\sum_{i=1}^{K} A_i \ldots\ldots (4)$$

where, ϒ represents the expected prediction accuracy of the classifier; *K* is the number of splits; and *A* is the accuracy of a fold (Arlot & Celisse, 2010; see Geisser,1975, for a detailed explanation of *K*-fold cross validation).

For each CDM-CR item, the initial score prediction model is built using extracted features along with an arbitrary subset of the training dataset. The partially trained *MLA* (i.e., developing model) is then iteratively refined using the remaining *K-1* training subsets. After the K^{th} iteration, the developed model is considered final for the automated score prediction process. For each CDM-CR item, three distinct score prediction models were developed, one each for SMO, Naïve Bayes, and J48 *MLAs*, and their expected accuracies and predictive consistency were then computed, compared, and contrasted.

Step Three – Automated Score Prediction

The developed models were used for automated scoring of the corresponding CDM-CR responses using data from the spring 2013 exam administration (i.e., Dataset-II). The model built using English CDM-CR responses were used to score English CDM-CR responses. Similarly, the model built using French CDM-CR responses were used to score French CDM-CR responses. The sample size for French ($n_{fr.}$) CDM-CRs is noticeably smaller than the English responses (see Table 1 for item-wise sample size details). In an attempt to minimize the impact of small sample size for the French training responses ($144 \leq n_{fr.} \leq 254$ per CDM-CR item), French responses were also machine translated into English using Google Translate™ (http://translate.google.ca), and then scored using the English CDM-CR model. Google Translate™ is a statistical machine-translation service which is freely available. It has also currently the most accurate service for translating text from French to English (Callison-Burch, 2009; Mehdad, Negri, & Federico, 2010).

Analysis of Automated Scoring Process

To evaluate the accuracy of the automated scoring process, a comprehensive evaluation scheme proposed by Williamson, Xi, and Breyer (2012) was used. That is, the model predictive power for English, French, and Translated-French CDM-CR items were evaluated on the basis of exact-agreement percentage and Kappa coefficient values.

Table 2

Viera & Garrett (2005) Guidelines for Interpreting Agreement Consistency Index – Kappa (κ)

Kappa Value	Strength of Agreement
< 0.0	Less than chance agreement
0.01 - 0.20	Slightly agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement

Agreement percentage was computed based on exact matching between human and predicted scores. Then, Cohen's (1960) Kappa coefficient (κ) was used to summarize the consistency of the score prediction model. Kappa is a summary estimate that measures agreement beyond chance. A Kappa of 1.0 indicates perfect agreement, whereas a Kappa of 0.0 indicates agreement equivalent to chance (random assignment) only. For the purpose of this research, guidelines presented in Table 2 from Viera and Garrett (2005) were used for interpreting κ statistics. The results are presented next.

Chapter 4: Results

The results are presented in three parts. Part one contains the results for English CDM-CR questions. Part two includes the results for French CDM-CR questions. Part three provides the results for translated-French (into English) CDM-CR question responses. In each part, the expected and absolute agreement percentages, kappa coefficients, score distribution, and variance for human and machine scores are presented.

English CDM-CR Questions

Expected Accuracy and Consistency. The overall results of model expected prediction accuracy and kappa (κ) consistency indices across eight English CDM-CR questions, using three automated score prediction models, are presented in Table 3. For the Naïve Bayes scoring model, the expected accuracy ranged from 84.9% to 98.2%, with corresponding consistency coefficients from 0.66 to 0.96. Expected accuracy for the SMO scoring model ranged from 92.1% to 98.3%, with corresponding kappa from 0.82 to 0.96. For the J48 scoring model, the expected accuracy ranged from 89.4% to 98.2%, with corresponding kappa between 0.75and 0.92.

For item 1, 2, 5, and 8, the SMO scoring model is expected to produce the highest agreement with human scores, and the high kappa value (κ≥0.82) also suggests the machine and human scoring will be almost perfectly consistent. For item 3 and 4, the J48 scoring model is expected to produce the highest agreement with human scores, and the kappa value of ≥ 0.91 also suggests almost perfect agreement.

Table 3

Expected Accuracy and Consistency for English CDM-CR questions Using K-Fold Cross-Validation Method

		Expected Accuracy and Consistency		
CDM-CR Question	**Count of Training Responses**	Naïve Bayes	SMO	J48
1	1266	*92.9%* ***0.77***	*97.5%* ***0.91***	*96.2%* ***0.88***
2	1259	*96.5%* ***0.92***	*98.3%* ***0.96***	*97.8%* ***0.95***
3	1255	*93.4%* ***0.84***	*95.6%* ***0.90***	*96.0%* ***0.91***
4	1269	*96.1%* ***0.85***	*97.9%* ***0.91***	*98.2%* ***0.92***
5	1255	*95.0%* ***0.87***	*96.0%* ***0.90***	*95.5%* ***0.88***
6	1205	*93.7%* ***0.55***	*95.9%* ***0.65***	*95.9%* ***0.68***
7	827	*98.2%* ***0.96***	*97.9%* ***0.96***	*97.6%* ***0.95***
8	1241	*84.9%* ***0.66***	*92.1%* ***0.82***	*89.4%* ***0.75***

For item 7, the Naïve Bayes is expected to produce almost perfect agreement with human scores. For item 6, both SMO and J48 are expected to equitably predict human scores. However, the J48 is expected to produce relatively higher consistency with human scoring (κ = 0.68). Table 3 also suggested that the SMO and J48 would be comparable in predicting close to human scores. Apart from item 1 and item 8, the scoring models are generally comparable in their agreement percentage and kappa values.

Absolute Accuracy and Consistency. Model absolute prediction accuracy and consistency indices, across eight English CDM-CR items, are presented in Table 4. For the Naïve Bayes scoring model, the prediction accuracy ranged from 83.6% to 98.7%, with corresponding kappas between 0.63and 0.97. The score prediction accuracy for the SMO scoring model ranged from 90.6% to 98.2%, with corresponding kappas between 0.79 and 0.96. Prediction accuracy for the J48 scoring model ranged from 88.1% to 98.2%, with corresponding kappas between 0.73and 0.95. Across all three scoring models, item 8 shared the lower bound of agreement with human scores. For this item, the best agreement observed was 90.6% with a kappa of 0.79.

Table 4

Absolute Accuracy and Consistency for English CDM-CR Questions

		Absolute Accuracy and Consistency		
CDM-CR Question	**Count of Prediction Responses**	Naïve Bayes	SMO	J48
1	856	*91.1%* ***0.72***	*97.3%* ***0.91***	*94.9%* ***0.83***
2	1268	*98.7%* ***0.97***	*98.2%* ***0.96***	*97.7%* ***0.95***
3	859	*91.9%* ***0.82***	*94.6%* ***0.88***	*94.1%* ***0.87***
4	1256	*96.9%* ***0.90***	*98.1%* ***0.94***	*97.7%* ***0.92***
5	1256	*96.6%* ***0.91***	*97.9%* ***0.94***	*98.2%* ***0.95***
6	1272	*93.9%* ***0.50***	*95.0%* ***0.57***	*95.1%* ***0.67***
7	848	*98.6%* ***0.97***	*98.0%* ***0.96***	*96.6%* ***0.93***
8	1260	*83.6%* ***0.63***	*90.6%* ***0.79***	*88.1%* ***0.73***

Based on the accuracy and consistency indices, the Naïve Bayes scoring models performed best in predicting scores for item 2 and 7. For item 1, 3, 4, and 8, the SMO scoring models best predicted the human scores. The J48 scoring model best predicted scores for item 5 and 6. Apart from item 8, the best score predictors are in agreement with human raters for at least 95% of the time and, often, the kappa value suggested perfect consistency (κ ≥ 0.81) with human scoring. The kappa values are graphically represented in Figure 5.

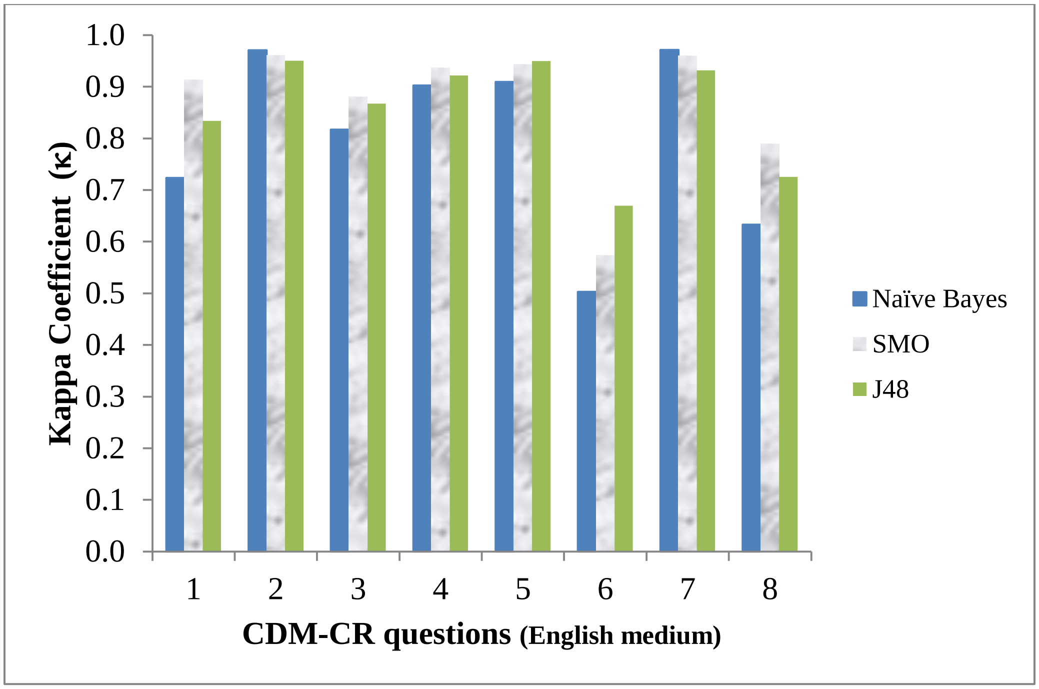


Figure 5. Kappa coefficients across English CDM-CR questions for three score prediction models.

Score Distribution and Variance. Score distributions for human and machine scoring are presented in Table 5. The corresponding variances are presented in Table 6. For all eight English CDM-CR questions, the machine score distributions are comparable to the human-produced score distributions. Moreover, the magnitude of the variances suggests that the machine-predicted scores does not have any systematic differences and are comparable to the aggregate human scores. Each variance value is identical across the three scoring model with only rounding error to differentiate the values. Specifically, item 6 had the lowest variability and item 7 had the highest variability (σ^2 = 0.25) in score distributions for human and predicted scores.

Table 5

Human and Computer Score Distributions for English CDM Responses

		Score Distributions							
		Zero				**One**			
CDM-CR Question	**Count of Prediction Responses**	Human	Naïve Bayes	SMO	J48	Human	Naïve Bayes	SMO	J48
1	856	*161* ***18.8%***	*183* ***21.4%***	*170* ***19.9%***	*167* ***19.5%***	*695* ***81.2%***	*673* ***78.6%***	*686* ***80.1%***	*689* ***80.5%***
2	1268	*447* ***35.3%***	*462* ***36.4%***	*468* ***36.9%***	*474* ***37.4%***	*821* ***64.7%***	*806* ***63.6%***	*800* ***63.1%***	*794* ***62.6%***
3	859	*278* ***32.4%***	*296* ***34.5%***	*302* ***35.2%***	*301* ***35.0%***	*581* ***67.6%***	*563* ***65.5%***	*557* ***64.8%***	*558* ***65.0%***
4	1256	*239* ***19.0%***	*262* ***20.9%***	*227* ***18.1%***	*216* ***17.2%***	*1017* ***81.0%***	*994* ***79.1%***	*1029* ***81.9%***	*1040* ***82.8%***
5	1256	*301* ***24.0%***	*332* ***26.4%***	*305* ***24.3%***	*304* ***24.2%***	*955* ***76.0%***	*924* ***73.6%***	*951* ***75.7%***	*952* ***75.8%***
6	1272	*92* ***7.2%***	*76* ***6.0%***	*68* ***5.3%***	*112* ***8.8%***	*1180* ***92.8%***	*1196* ***94.0%***	*1204* ***94.7%***	*1160* ***91.2%***
7	848	*418* ***49.3%***	*414* ***48.8%***	*411* ***48.5%***	*411* ***48.5%***	*430* ***50.7%***	*434* ***51.2%***	*437* ***51.5%***	*437* ***51.5%***
8	1260	*410* ***32.5%***	*443* ***35.2%***	*443* ***35.2%***	*392* ***31.1%***	*850* ***67.5%***	*817* ***64.8%***	*817* ***64.8%***	*868* ***68.9%***

Table 6

Variability in Human and Machine scores for English CDM-CR Questions

CDM-CR Question	**Count of Prediction Responses**	Human	Naïve Bayes	SMO	J48
1	856	0.15	0.17	0.16	0.16
2	1268	0.23	0.23	0.23	0.23
3	859	0.22	0.23	0.23	0.23
4	1256	0.15	0.17	0.15	0.14
5	1256	0.18	0.20	0.18	0.18
6	1272	0.07	0.06	0.05	0.08
7	848	0.25	0.25	0.25	0.25
8	1260	0.22	0.23	0.23	0.21

French CDM-CR Questions

Expected Accuracy and Consistency. The overall expected predictive accuracy and consistency for French CDM-CR models are presented in Table 7. Across eight CDM-CR items, expected accuracy for the Naïve Bayes scoring models ranged from 81.8% to 97.9%, with corresponding kappas between 0.50 and 0.89. Expected accuracy for the SMO scoring models ranged between 92.1% and 99.2%, with corresponding kappas from 0.78 to 0.97. For the J48 scoring models, the expected accuracy ranged from 89.1% to 99.1%, with corresponding kappas between 0.54 and 0.90.

For item 1, the Naïve Bayes is expected to produce the highest agreement with human scores. For item 2, the expected accuracy and consistency are equal across all three scoring models. For item 4 and item 5, the J48 and SMO scoring models are expected to agree with human 99% of the time, respectively. For item 3, 6, and 8 the SMO scoring model is expected to produce higher agreement with human scores. However, the kappa value of 0.20 for item 6 suggested the predicted scores may not be consistent with human-produced scores. For item 7, the J48 scoring model is expected to best predict the human scores and the kappa value of 0.89 also suggests the machine and human scoring to be almost perfectly consistent.

Table 7

Expected Accuracy and Consistency for French CDM-CR questions Using K-Fold Cross-Validation Method

CDM-CR Question	**Count of Training Responses**	**Expected Accuracy and Consistency**		
		Naïve Bayes	SMO	J48
1	226	*95.1%* ***0.54***	*93.8%* ***0.43***	*90.7%* ***0.04***
2	237	*97.9%* ***0.89***	*97.9%* ***0.89***	*97.9%* ***0.89***
3	238	*92.9%* ***0.70***	*93.3%* ***0.73***	*89.1%* ***0.54***
4	226	*92.5%* ***0.45***	*98.2%* ***0.79***	*99.1%* ***0.90***
5	239	*93.7%* ***0.82***	*99.2%* ***0.97***	*98.7%* ***0.96***
6	254	*91.7%* ***0.04***	*94.5%* ***0.20***	*93.7%* ***0.17***
7	144	*93.1%* ***0.85***	*93.1%* ***0.85***	*94.4%* ***0.88***
8	242	*81.8%* ***0.50***	*92.1%* ***0.78***	*92.1%* ***0.77***

Absolute Accuracy and Consistency. The absolute predictive accuracy and consistency indices, across eight French CDM-CR items, are presented in Table 8. For the Naïve Bayes scoring model, the predictive accuracy ranged from

75.0% to 97.1%, with corresponding kappas from 0.29 to 0.86. Absolute prediction accuracy for the SMO scoring models ranged from 87.3% to 98.8%, with corresponding kappas from 0.61 to 0.90. For J48 scoring models, the prediction accuracy ranged from 88.2% and 98.8%, with corresponding kappas from 0.62 to 0.90.

Table 8

Absolute Accuracy and Consistency for French CDM-CR Questions

CDM-CR Question	**Count of Prediction Responses**	**Absolute Accuracy and Consistency**		
		Naïve Bayes	SMO	J48
1	149	*91.3%* ***0.40***	*91.3%* ***0.47***	*89.9%* ***0.18***
2	241	*97.1%* ***0.86***	*97.1%* ***0.86***	*97.1%* ***0.86***
3	143	*90.9%* ***0.63***	*95.8%* ***0.84***	*95.1%* ***0.82***
4	246	*96.7%* ***0.77***	*98.8%* ***0.90***	*98.8%* ***0.90***
5	246	*92.3%* ***0.78***	*98.4%* ***0.95***	*98.4%* ***0.95***
6	241	*97.1%* ***-0.01***	*96.3%* ***0.38***	*97.5%* ***0.39***
7	156	*96.2%* ***0.91***	*97.4%* ***0.94***	*96.8%* ***0.92***
8	228	*75.0%* ***0.29***	*87.3%* ***0.61***	*88.2%* ***0.62***

For item 2, there was a performance tie across all scoring models with 97.1% agreement and almost perfect consistency of 0.86. Another performance tie exists, between the SMO and J48 scoring models, for item 4 and item 5. Across both scoring models, item 4 had the accuracy rate of 98.8% with almost perfect consistency of 0.90 and item 5 had an accuracy of 98.4% with almost perfect consistency of 0.95. For item 1, 3, and 7, the SMO scoring models was best in predicting the human scores. The J48 scoring model best predicted scores for item 6 and 8. Apart from item 1 and 8, the best score predictors are in agreement with humans for at least 96% of the time, and the kappa value (κ ≥ 0.81) suggests a high level of agreement with human-produced scores. Across all three scoring models, item 8 yields the lower bound of prediction accuracies, in which case the maximum agreement was observed to be 88.2% with the corresponding kappa value of 0.62. The kappa values across eight CDM-CR questions are graphically represented in Figure 6.

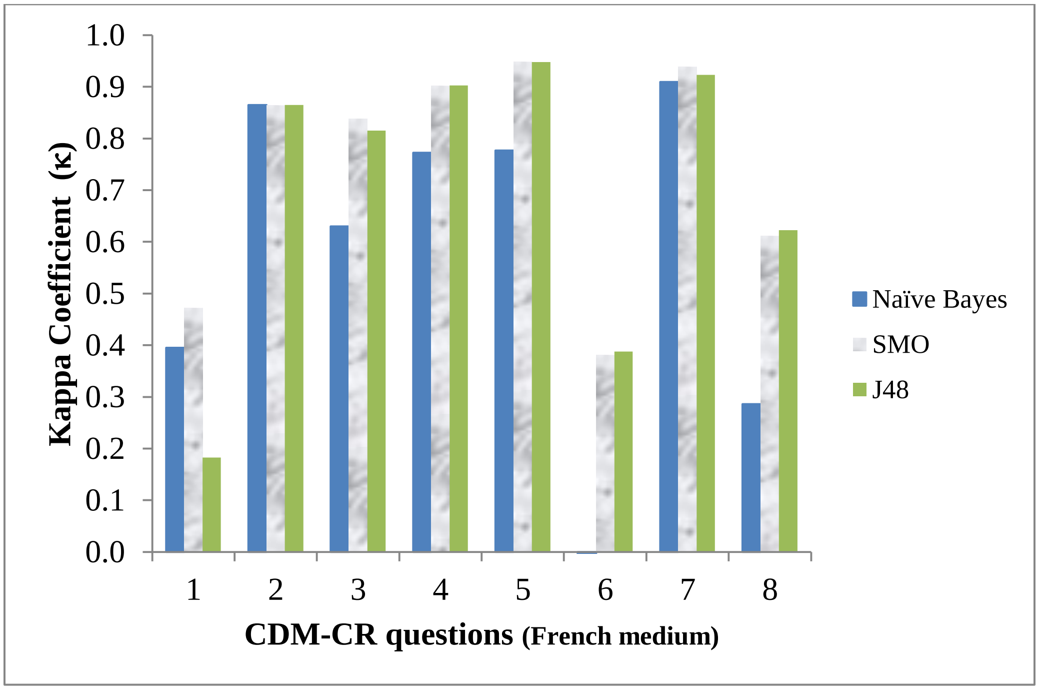


Figure 6. Kappa coefficients across French CDM-CR questions for three score prediction models.

Score Distribution and Variance. Table 9 contains the distributions for human-produced and machine-predicted scores. The corresponding variances are presented in Table 10. For item 1, 4, and 6 the lower variability in predicted scores could be due to the small number of responses for these items (≤ 15 CDM responses with zero score) for model training process. For item 8, score distributions for the J48 model deviated by 10% from the human score distribution. Apart from problem associated with small sample size, the predicted score distributions for French CDM-CR does not shown any systematic difference and are comparable with the human score distributions.

Table 9

Human and Computer Score Distributions for French CDM Responses

CDM-CR Question	**Count of Prediction Responses**	**Score Distributions**							
		Zero				**One**			
		Human	Naïve Bayes	SMO	J48	Human	Naïve Bayes	SMO	J48
1	149	*16* ***10.7%***	*7* ***4.7%***	*11* ***7.4%***	*3* ***2.0%***	*133* ***89.3%***	*142* ***95.3%***	*138* ***92.6%***	*146* ***98.0%***
2	241	*28* ***11.6%***	*31* ***12.9%***	*31* ***12.9%***	*31* ***12.9%***	*213* ***88.4%***	*210* ***87.1%***	*210* ***87.1%***	*210* ***87.1%***
3	143	*22* ***15.4%***	*19* ***13.3%***	*22* ***15.4%***	*23* ***16.1%***	*121* ***84.6%***	*124* ***86.7%***	*121* ***84.6%***	*120* ***83.9%***
4	246	*15* ***6.1%***	*23* ***9.3%***	*18* ***7.3%***	*18* ***7.3%***	*231* ***93.9%***	*223* ***90.7%***	*228* ***92.7%***	*228* ***92.7%***
5	246	*48* ***19.5%***	*61* ***24.8%***	*50* ***20.3%***	*48* ***19.5%***	*198* ***80.5%***	*185* ***75.2%***	*196* ***79.7%***	*198* ***80.5%***
6	241	*6* ***2.5%***	*1* ***0.4%***	*9* ***3.7%***	*4* ***1.7%***	*235* ***97.5%***	*240* ***99.6%***	*232* ***96.3%***	*237* ***98.3%***
7	156	*47* ***30.1%***	*49* ***31.4%***	*47* ***30.1%***	*46* ***29.5%***	*109* ***69.9%***	*107* ***68.6%***	*109* ***69.9%***	*110* ***70.5%***
8	228	*55* ***24.1%***	*48* ***21.1%***	*38* ***16.7%***	*32* ***14.0%***	*173* ***75.9%***	*180* ***78.9%***	*190* ***83.3%***	*196* ***86.0%***

Table 10

Variability in Human and Machine scores for French CDM-CR Questions

CDM-CR Question	**Count of Prediction Responses**	Human	Naïve Bayes	SMO	J48
1	149	0.10	0.05	0.07	0.02
2	241	0.10	0.11	0.11	0.11
3	143	0.13	0.12	0.13	0.14
4	246	0.06	0.09	0.07	0.07
5	246	0.16	0.19	0.16	0.16
6	241	0.02	0.004	0.04	0.02
7	156	0.21	0.22	0.21	0.21
8	228	0.18	0.17	0.14	0.12

Translated-French CDM-CR Questions

Absolute Accuracy and Consistency–A Comparison. Responses for French CDM-CR questions were first translated into English and then scored using prediction models of English CDM-CR questions. The intent was to improve the predictive accuracy and consistency for the French CDM questions, which had relatively low sample sizes (≤ 246) across the eight CDM questions. The results for the absolute predictive accuracy and consistency indices, across eight translated-French CDM-CR items, are shown in Table 11. The gain in predictive accuracy and consistency between French and translated-French CDM-CR items is summarized in Table 12.

Table 11

Absolute Accuracy and Consistency for Translated-French CDM Responses Using Score Prediction Models from English CDM-CR Questions

CDM-CR Question	**Count of Prediction Responses**	**Absolute Accuracy and Consistency**		
		Naïve Bayes	SMO	J48
1	149	*86.6%* ***0.55***	*88.6%* ***0.59***	*87.9%* ***0.54***
2	241	*98.8%* ***0.94***	*98.8%* ***0.94***	*99.2%* ***0.96***
3	143	*91.6%* ***0.69***	*93.0%* ***0.74***	*94.4%* ***0.79***
4	246	*87.8%* ***0.45***	*99.6%* ***0.96***	*98.4%* ***0.87***
5	246	*91.9%* ***0.76***	*94.7%* ***0.84***	*94.7%* ***0.84***
6	241	*95.0%* ***0.23***	*98.3%* ***0.59***	*98.8%* ***0.76***
7	156	*92.3%* ***0.82***	*96.2%* ***0.91***	*96.8%* ***0.92***
8	228	*85.5%* ***0.62***	*89.0%* ***0.68***	*88.2%* ***0.65***

In Table 12, positive values suggest that the translation had improved the performance of the score prediction processes whereas negative values indicate a decrease in performance of the score prediction model. For example, with item 6, +2.1% for the SMO model suggested that the machine translation improved the agreement between human and machine-predicted scores by 2.1%. That is, the accuracy increased from 96.3% to 98.3%. Similarly, for item 6, -2.1% for Naïve Bayes model indicates that the translation process reduced the agreement between human and machine-predicted scores by 2.1%. That is, the accuracy decreased from 97.1% to 95%.

Table 12

Gain in Score Prediction Accuracy and Consistency between French and Translated-French CDM-CR Questions

CDM-CR Question	**Count of Prediction Responses**	**Absolute Accuracy and Consistency**		
		Naïve Bayes	SMO	J48
1	149	*-4.7%* ***0.15***	*-2.7%* ***0.12***	*-2.0%* ***0.36***
2	241	*1.7%* ***0.07***	*1.7%* ***0.07***	*2.1%* ***0.10***
3	143	*0.7%* ***0.06***	*-2.8%* ***-0.10***	*-0.7%* ***-0.03***
4	246	*-8.9%* ***-0.32***	*0.8%* ***0.06***	*-0.4%* ***-0.04***
5	246	*-0.4%* ***-0.02***	*-3.7%* ***-0.11***	*-3.7%* ***-0.11***
6	241	*-2.1%* ***0.23***	*2.1%* ***0.21***	*1.2%* ***0.38***
7	156	*-3.8%* ***-0.09***	*-1.3%* ***-0.03***	*0.0%* ***0.00***
8	228	*10.5%* ***0.33***	*1.8%* ***0.07***	*0.0%* ***0.03***

Apart from item 2, 6, and 8, most scoring models had shown a reduction or a negligible gain in accuracy and consistency. For item 1, the reduced agreement is paired with increased consistency of machine-predicted scores. The same conclusion is true for item 6 for the Naïve Bayes model only. Naïve Bayes also produced extreme results for item 4, and 8. For item 4, Naïve Bayes reduced the agreement by 9% and kappa by 0.32 points. For item 8, Naïve Bayes showed a gain in agreement by 11% and kappa by 0.33 points. The gain in accuracy and consistency across the three score prediction models is graphically represented in Figure 7 and 8, respectively.

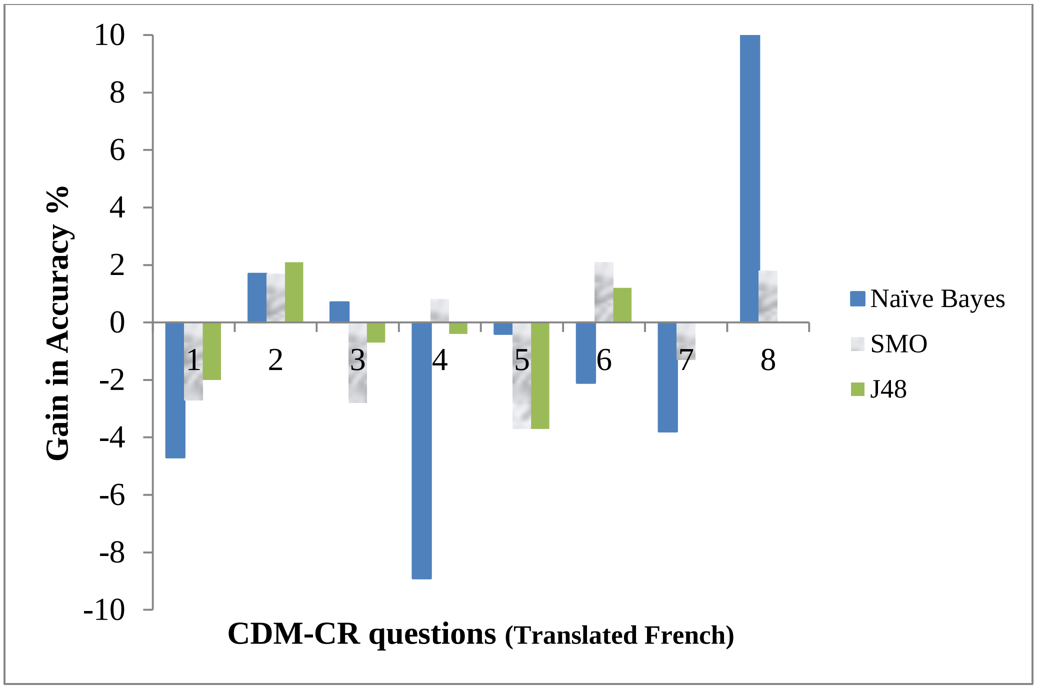


Figure 7. Gain in accuracy as a result of machine translation across eight CDM CR questions for three score prediction models.

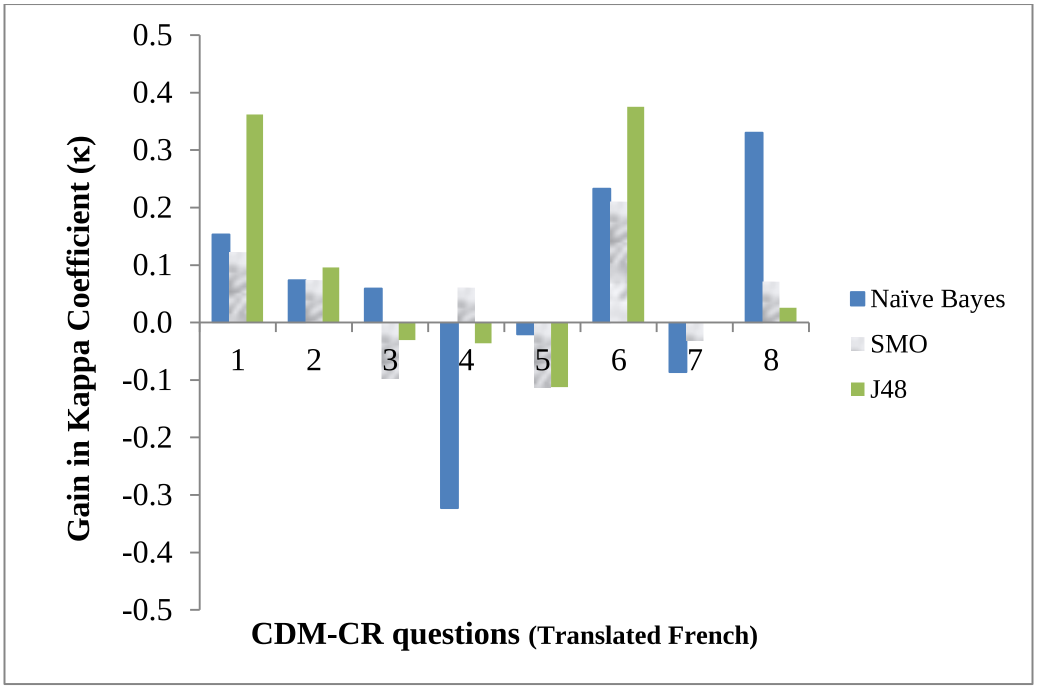
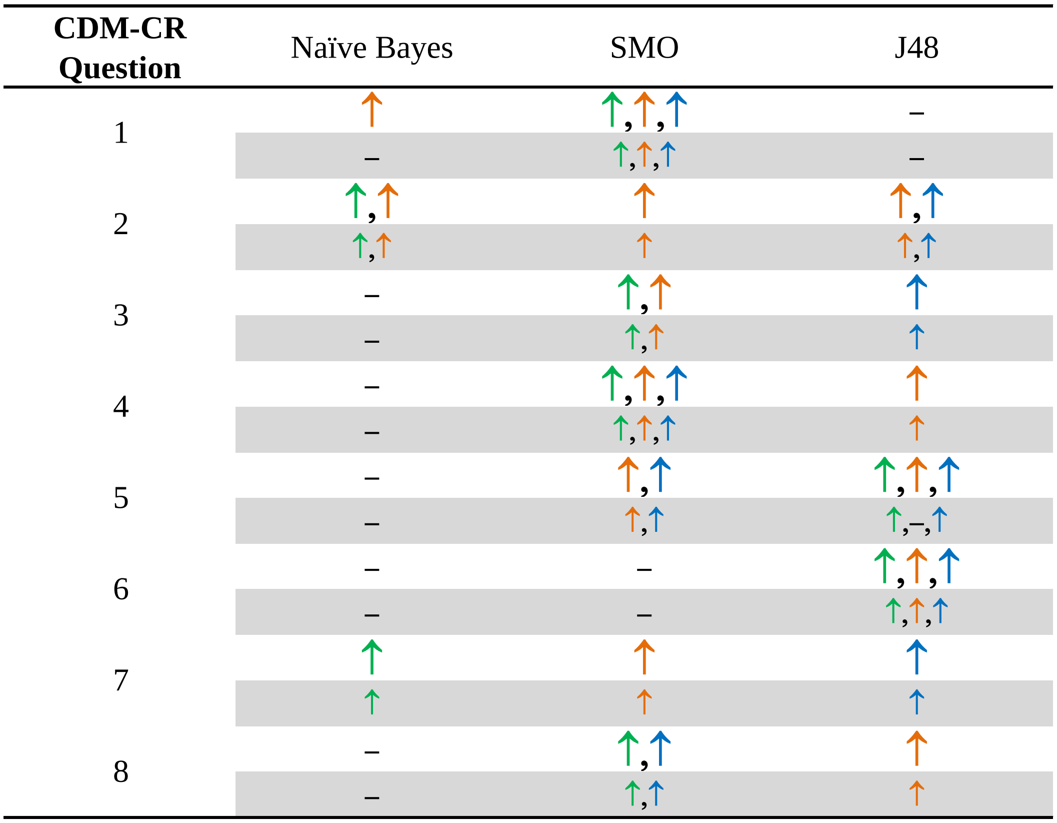


Figure 8. Gain in consistency as a result of machine translation across eight CDM-CR questions for three score prediction models.



CDM-CR Question	Naïve Bayes	SMO	J48
1	↑	↑,↑,↑	–
	–	↑,↑,↑	–
2	↑,↑	↑	↑,↑
	↑,↑	↑	↑,↑
3	–	↑,↑	↑
	–	↑,↑	↑
4	–	↑,↑,↑	↑
	–	↑,↑,↑	↑
5	–	↑,↑	↑,↑,↑
	–	↑,↑	↑,–,↑
6	–	–	↑,↑,↑
	–	–	↑,↑,↑
7	↑	↑	↑
	↑	↑	↑
8	–	↑,↑	↑
	–	↑,↑	↑

Legends:

↑	Best accuracy for English CDM-questions
↑	Best accuracy for French CDM-questions
↑	Best accuracy for Translated-French CDM-questions
▨	Gray cells showing best consistency (Kappa) measure for CDM-questions

Figure 9. Overall performance for three scoring framework across eight CDM questions.

The absolute performance of three scoring algorithm across English, French, and Translated-French CDM responses is summarized in Figure 9. Overall, the scoring framework developed using SMO produced the best accuracy and consistency results for predicting human scores, followed by the automatic scoring using J48 learning algorithm. Item 2 is the only item which has the performance tie across all three MLAs for French CDM responses. Naïve Bayes performed well only for three CDM questions, and it did not competed in scoring for any Translated-French CDM item. Apart from item 1, J48 program generally performed well, and often times there were performance tie between SMO and J48, suggesting that they are generally comparable.

In summary, the agreement percentages and kappa coefficients were comparable for the automated score prediction of the eight CDM-CR writing prompts. The concordance between expected and absolute accuracy and consistency indices was generally in alignment. Although the score prediction models for English and French CDM-CR questions were comparable, English CDM-CR scoring models had relatively better performance than French score prediction models. But scoring of the translated-French CDM-CR questions did show performance gain for three of the eight CDM-CR items. The discussion and conclusion are presented next.

Chapter 5: Discussion

Summary of Findings Given Purposes of the Study

Clinical-decision making (CDM) questions are regarded as an integral part of the medical licensing examination process because this item type deals with skills required to think, reason, and solve medical problems. Such a written-response format carries little or no subjective interpretation and, as a result, has the advantage of being scored using natural language processing techniques. The purpose of this study was to demonstrate and evaluate the strength of three supervised machine learning algorithms (MLAs) for developing the automated scoring framework using open-source machine learning environment. I also investigated the feasibility of employing machine-translation in the automated scoring process.

The present study was designed to address three different purposes as presented in the introduction chapter. Each research purpose is followed by a description of how that purpose was answered or addressed is the present study.

1) To develop and demonstrate the automated scoring framework using open-source technologies for a multilingual medical licensing context: Three supervised machine learning algorithms were used to develop and demonstrate the automated scoring framework. The performance of the algorithms were then evaluated using two independent performance measures, i.e., accuracy and consistency. The automated scoring framework was developed for English, French, and translated-French CDM-CR questions. The elements of the developed framework were presented in Figure 4. The machine learning algorithms included Naïve Bayes – probabilistic learning, Sequential Minimum Optimization (SMO) – support vector learning, and J48 program – decision-tree learning. Two performance measures, exact agreement percentage (accuracy) and un-weighted Kappa (consistency), were computed between human and machine scoring.

2) To compare the strength of multilingual score prediction engines with scores obtained from human raters: The results from the present study helped demonstrate that the open-source scoring framework works extremely well for eight medical CDM question prompts, with average maximum prediction accuracies (in relation to human raters) of 96.4%, 95.6%, and 95.1%, for English, French, and Translated-French CDM responses, respectively. The variances of the score distributions (see Table 6 and Table 10) between human-produced and machine-predicted scores were also found to be comparable. For example, the score distributions for English CDM responses (Table 6) across the three scoring frameworks are almost identical to the human score distributions with only slight difference between the human and machine ratings (≤ 0.02). Similarly, the predicted score distributions for French CDM responses (Table 10) is comparable to human-produced score distributions, with only negligible difference between them (≤ 0.03) suggesting an overall strength, flexibility, and employability of open-source automated scoring for multilingual medical licensing examinations. In sum, the results from this study suggest that integrating technological innovations for high-stake assessment situations is feasible and, from an economic point-of-view, highly desirable.

3) To contextualize machine-translation for raising the performance gain when restricted sample size counters the model building process: Both positive and negative gain in accuracy and consistency was observed when machine-translation (Google Translate™) was used as a mechanism for dealing with a small sample size of French CDM questions. It was reported that the machine-translation raised the gain in the consistency measure but negatively influenced the agreement accuracy. However, there are some instances which showed positive gain on both measures. For example, Naïve Bayes for item 8 had a positive gain in accuracy by 10.5% and also raised the consistency by 0.33. A similar pattern (but with less magnitude) was also observed for items 2 and 6. Conversely, for some translated-French item responses, the accuracy decreased while the consistency increased, which suggests a non-linear gain from the machine-translation process.

To date, Google Translate™ has been demonstrated to be the most accurate application for translating text from French to English (Callison-Burch, 2009; Mehdad, Negri, & Federico, 2010). The results from this study suggest that such contextualization (an attempt to improve predictive accuracy and consistency by using a model from a large alternate training sample— English) may not always result in a performance gain. However, scoring systems supplemented with real-time machine-translation sub-system exists (see Higgins, 2013; Pérez et al., 2005) and assessments scientists might consider developing systems that are capable of automatically translating examinees' responses in the alternate language.

Conclusion and General Discussion

The performance measures across the three MLAs demonstrated that while their prediction accuracy and consistency are promising, they are not identical. That is, the performance ranking for the three MLAs differs across the eight CDM questions. For example, Naïve Bayes had the lowest prediction accuracy for item 1 using the English CDM-CR (91.1%), but for item 2 this MLA had the highest prediction accuracy (98.7%) when compared to SMO and J48. Clearly, the order of the performance measures are not consistent across the MLAs and there is reason to believe that identifying a high performing MLA in advance could reduce overhead for developing the best performing scoring framework.

But selecting the best performing scoring framework is not straight forward, especially when competing varieties of learning algorithms (e.g., probabilistic, decision-tree, neural network, etc.) are available. A unified technique or method does not exists that could reveal, in advance, which MLA would work best for a given classification problem (Hastie, Tibshirani, & Friedman, 2009). Empirical comparisons have shown that the best machine learning algorithm varies from application to application (Domingos, 2012). These differences occur because the prediction models which performs best in one content area may occasionally perform poorly in a different content area. Conversely, the prediction models with poor average performance in one application can occasionally perform exceptionally well in a different application (Caruana & Niculescu-Mizil, 2006).

The literature in the machine learning domain does not clearly demarcate which MLA should be used given the content and/or attribute of the dataset. However, some studies suggest that with support vector learning, one should expect high predictive accuracy (i.e., low error rates) when learning is run on data in its basic forms and when a limited dataset is available for model training (Bradski & Kaehler, 2008) because support vector learning is capable of making good decisions about data points that were not been supplied as part of training dataset (Harrington, 2012). This phenomenon is often called prediction generalization (Domingos, 2012). On the other hand, decision-tree learning is efficient because it is considered to be a simple and flexible approach to classification using the so called information gain approach (Costa et al., 2013; Domingos, 2012). Decision trees are also commended for scalability, speed, immunity towards outliers in training dataset, robustness to missing data, and ease in interpretability of the classification models (Harrington, 2012; Hastie, Tibshirani, & Friedman, 2009). However, for small training samples (*n*< 1000), decision tree learning seldom provides predictive accuracy that could be achieved using other prediction algorithms, such as Naïve Bayes (Hastie, Tibshirani, & Friedman, 2009; Domingos, 2012). Naïve Bayes learning, which uses probability theory for text classification, is conceptually simple, efficient in practice, and considered to be an optimal classification strategy is most situations when small amounts of training data are available (Domingos, 2012; Manning, Raghavan, & Schütze, 2008; Zhang, 2004).

It is unrealistic to assume that a human marker would have all of the pre-knowledge about the essays during standard setting and/or the marking process and, for such responses, the marker could acquire a peer judgment. However, the machine does not have this freedom. While building the automated scoring model it is assumed that the training essays provided to the learning algorithm would statistically resembles the essays instances encountered later on during the score prediction process. This is an idealistic view in any real-world situation (Dekel, Shamir, & Xiao, 2010). One way to deal with this challenge is to employ unsupervised (or semi-supervised) algorithms for developing score prediction frameworks.

There are many consideration required when making decisions about selecting the best scoring framework (i.e., the underlying classification algorithm). One way to make this decision in advance is to run the cross-validation for multiple MLA and select the MLA based on the best expected accuracy from cross-validation (Hastie, Tibshirani, & Friedman, 2009). However, in the present study, the expected accuracy by means of cross-validation was not always consistent with the absolute predictive accuracy of the model. For example, as shown in Table 3 and 4, the cross-validation results for item 5 suggest the expected highest prediction accuracy for SMO (expected accuracy = 96% and κ = 0.90), but the absolute prediction performance reveals that the J48 had the best predictive accuracy (absolute accuracy = 98.2% and κ= 0.95). The opposite is true for item 3. In fact, for some CDM questions the expected and absolute prediction accuracies are consistent and also identical in some cases across three MLAs.

Performance patterns for MLAs across three linguistic conditions (Figure 9) suggested that the support vector learning (SMO) is a promising prediction technique for almost all CDM questions. Decision-tree learning (J48) was the second best predictor. Probabilistic learning (Naïve Bayes) was the third best predictor. However, the third best predictor — Naïve Bayes, remained close to the best performing predictors by about 6% (apart from items 4 and 8) and was shown to be the top performer for item 2 and item 7 for the English CDM questions, and item 1 for the French CDM question. SMO and J48 were comparable because, for the most part, their absolute prediction accuracies are similar and they deviated only by 1.4% or less across the study conditions. This outcome suggests that the automated scoring framework developed using SMO and J48 are comparable in predicting human scores followed by the Naïve Bayes score prediction.

In sum, the proposed automated scoring frameworks are all quite consistent with one another and with human-produced scores. However, a policy still needs to be outlined as to what is an acceptable outcome for each scoring framework and how the scoring result would be used in the high-stakes medical licensing testing situation. The proposed framework could be used in two ways. First, for the existing two human marking structure (at the licensing authority), where scoring judgement are sought from two human markers, one human marker could be replaced with an automated scoring engine and the second human judgement should only be sought in cases where there is a discrepancy between human and machine scores. Second, to use the scoring framework as an additional quality assurance indicator to confirm the scores assign by the human markers are both consistent and accurate. The former implication is a more obvious resource-saver than the later from the point-of-view of money and time.

Limitations of the Study

The agreement among human scorers was not available, and thus the predictive accuracy achieved in this study could not be compared with human-human agreements. Further, the feature extraction mechanism discussed in this study was based on restricted set of features, as only limited feature extraction options are available in the open-source machine learning environment. That is, the semantic and/or linguistics characteristics of essays were determined by the co-occurrence of n-grams in the training dataset. This limits the employability of proposed framework in assessment situations where examinees are credited for their style and organization of writings.

Recommended Direction for Future Research

Three future directions for research are suggested. First, the agreement between two human markers for the medical CDM questions is currently not available. Hence, the accuracy of the human-rated CDMs is currently unknown. Future studies should be conducted to evaluate the agreement rate from CDM-CR scoring, between human markers, and with machine scoring. Second, the biannual licensing examination carries a number of new CDM questions that could not be scored with the framework proposed in this study. Thus the automated scoring framework capable of scoring both reused and new CDM questions is highly desirable. Future studies may work to extend the existing framework by incorporating the unsupervised (or semi-supervised) machine learning algorithms and evaluating its prediction strengths. Third, most written assessments grant credit to examinees for *style of writing*, *grammatical accuracy*, and/or *organization of ideas*. Such feature, also called micro-features (Bridgeman, Trapani, & Attali, 2012), requires extraction procedures beyond *n*-grams (Baayen, Hendrix, & Ramscar, 2013). Further research is required to investigate the micro-feature extraction techniques that could be incorporated in the proposed scoring framework. In sum, different assessment circumstances should be replicated for generalizing the strength of using open-source machine learning environment.

References

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40-79.

Attali, Y. (2013), Validity and Reliability of Automated Essay Scoring. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 124-135). New York: Psychology Press.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125-141.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the Combinatorial Explosion: An Explanation of n-gram Frequency Effects Based on Naïve Discriminative Learning. *Language and Speech, 56(3) 329–347*.

Bradski, G., & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. Sebastopol, CA: O'reilly.

Brew, C., & Leacocle, C. (2013). Automated Short Answer Scoring.In M.D. Shermis& J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 136-152). New York: Psychology Press.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40.

Callison-Burch, C. (2009, August). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 286-295). Association for Computational Linguistics.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168). ACM.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3), 5432-5435.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Costa, K., Ribeiro, P., Camargo, A., Rossi, V., Martins, H., Neves, M., ... & Papa, J. P. (2013, August). Comparison of the techniques decision tree and MLP for data mining in SPAMs detection to computer networks. In *Innovative Computing Technology (INTECH), 2013 Third International Conference on* (pp. 344-348). IEEE.

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international conference on World Wide Web* (pp. 519-528). ACM.

De, A., & Kopparapu, S. K. (2011, September). An unsupervised approach to automated selection of good essays. In *Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE* (pp. 662-666). IEEE.

Dekel, O., Shamir, O., & Xiao, L. (2010). Learning to classify with missing and corrupted features. *Machine learning*, *81*(2), 149-178.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87.

Elliot, N., & Klobucar, A. (2013). 2 Automated Essay Evaluation and the Teaching of Writing. In M.D. Shermis& J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 16-35). New York: Psychology Press.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2005). Weka. In *Data Mining and Knowledge Discovery Handbook* (pp. 1305-1314). Springer US.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, *20*(15), 2479-2481.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, *3*(1998), 1-10.

Geisser, S. (1975). The predictive sample reuse method with applications.*Journal of the American Statistical Association*, *70*(350), 320-328.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18.

Harrington, P. (2012). *Machine Learning in Action*. Greenwich, CT: Manning Publications Co..

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York: Springer.

Higgins, D. (2013, April). *ETS participation in the ASAP short answer scoring challenge*. Paper presentation at the annual meeting of the National Council on Measurement in Education, San Franscisco, CA.

Keith, T. Z. (2003). Validity of automated essay scoring systems. In M.D. Shermis & J.C. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (pp. 147-167). Mahwah, New Jersey: Lawrence Erlbaum associates, Inc..

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint Conference on artificial intelligence* (Vol. 14, pp. 1137-1145). Lawrence Erlbaum Associates Ltd.

Kononenko, I. (1990). Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition. In Weileinga, B., ed., *Current trends in knowledge acquisition.*IOS Press.

Lee, C. H., & Yang, H. C. (2009). Construction of supervised and unsupervised learning systems for multilingual text categorization. *Expert Systems with Applications*, *36*(2), 2400-2410.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge University Press.

Mayfield, E., & Rose, C. P. (2013), LightSIDE: Open Source Machine Learning for Text. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 124-135). New York: Psychology Press.

Mayfield, E., Adamson, D., & Rose, C. P. (2013). *Researcher's User Manual*. Retrieved from http://lightsidelabs.com/.

Mehdad, Y., Negri, M., & Federico, M. (2010, June). Towards cross-lingual textual entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for*

Computational Linguistics (pp. 321-324). Association for Computational Linguistics.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Myers, M. (2003). What can computers and AES contribute to a K-12 writing program?. In M.D. Shermis & J.C. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (pp. 1-20). Mahwah, New Jersey: Lawrence Erlbaum associates, Inc..

Page, E. B. (1966). The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, *47*(5), 238-243.

Pazzani, M.J. (1996). Search for dependencies in Bayesian classifiers. In Fisher, D., and Lenz, H.J., eds., *Learning from data: artificial intelligence and statisitics V.* Springer Verlag.

Pérez, D., Alfonseca, E., Rodríguez, P., Gliozzo, A., Strapparava, C., & Magnini, B. (2005). About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. *International Journal Signos,*38(59), 325-343.

Pitler, E., Louis, A., & Nenkova, A. (2009, August). Automatic sense prediction for implicit discourse relations in text. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 683-691). Association for Computational Linguistics.

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Retrived from: http://research.microsoft.com/pubs/69644/tr-98-14.pdf

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning* (Vol. 1). San Mateo, CA: Morgan Kaufmann Publishers.

Rich, C. S., Schneider, M. C., & D'Brot, J. M. (2013). Applications of Automated Essay Evaluation in West Virginia. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 99-123). New York: Psychology Press.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(3), 569-575.

Rudner, L.M. & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *Journal of Technology, Learning, and Assessment*, 1 (2). Retrieved from http://www.jtla.org.

Schultz, M. T. (2013). The IntelliMetric™ Automated Essay Scoring Engine—A Review and an Application to Chinese Essay Scoring. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 89-98). New York: Psychology Press.

Shermis, M.D., & Burstein, J.C. (2003). Introduction. In M.D. Shermis& J.C. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (pp. 1-20). Mahwah, New Jersey: Lawrence Erlbaum associates, Inc..

Shermis, M.D., & Hamner, B. (2012). *Contrasting state-of-the-art automated scoring of essays: analysis.* Paper presented at the National Council on Measurement in Education, Vancouver, BC, Canada.

Shermis, M.D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. In M.D. Shermis& J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 313-346). New York: Psychology Press.

Souza, C. R. (2010, March 17). Kernel Functions for Machine Learning Applications. *Science, computing and machine learning.* http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*,*2*, 45-66.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, *37*(5), 360-363.

Weigle, S. C. (2013). English as a Second Language Writing and Automated Essay Evaluation. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 36-54). New York: Psychology Press.

Williamson, D. M. (2013). 10 Probable Cause Developing Warrants for Automated Scoring of Essays. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: current application and new directions* (pp. 153-180). New York: Psychology Press.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2-13.

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations.

Xiong, W., Song, M., & deVersterre, L. (2012). A Comparative Study of an Unsupervised Word Sense Disambiguation Approach. In P. McCarthy, & C. Boonthum-Denecke (Eds.) *Applied Natural Language Processing: Identification, Investigation and Resolution* (pp. 412-422). Hershey, PA: Information Science Reference. doi:10.4018/978-1-60960-741-8.ch024

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011, June). A New Dataset and Method for Automatically Grading ESOL Texts. In *ACL* (pp. 180-189).

Zhang, H. (2004). The optimality of Naïve Bayes. In *Proceedings of the FLAIRS Conference* (Vol. 1, No. 2, pp. 3-9).