Text classification of student predicate use for automatic misconception categorization

1st Brian A. Landrón-Rivera
Electrical and Computer Engineering
University of Puerto Rico Mayagüez Campus
Mayagüez, P.R.
brian.landron@upr.edu

2nd Nayda G. Santiago
Electrical and Computer Engineering
University of Puerto Rico Mayagüez Campus
Mayagüez, P.R.
nayda.santiago@ece.uprm.edu

3rd Aidsa Santiago
Engineering
University of Puerto Rico Mayagüez Campus
Mayagüez, P.R.
aidsa.santiago@upr.edu

4th J. Fernando Vega-Riveros
Electrical and Computer Engineering
University of Puerto Rico Mayagüez Campus
Mayagüez, P.R.
aidsa.santiago@upr.edu

***Abstract*—This Research Category Full Paper presents an approach for categorizing student misconceptions about dynamics and heat transfer using text classification. Research in educational engineering describes how science concepts can be ontologically categorized into two major categories (substances and processes) according to their nature. Students can acquire misconceptions by incorrectly categorizing concepts. Predicate tests help to determine when misconceptions have occurred and aid in customizing curriculum content. However predicate tests rely on time-consuming, manual labor. The main goal of this research was to show how predicate tests could be automated with text classification models using a previously annotated dataset. We compared classifier performance between WEKA's Support Vector Machines, Multinomial Naïve Bayes, Logistic Regression, and Bayesian Logistic Regression implementations using the emergent process and sequential process ontological categories as labels. We compared model performance using WEKA's N-Gram tokenizer with 3-grams vs. using Java's WordTokenizer to convert our word dataset to numerical vectors. Models were evaluated using 10-fold cross-validation considering accuracy, F-measure, and kappa coefficient as measures of performance. We have shown the feasibility of using text classification for misconception assessment. Our implementation of predicate test automation can play an important role in speeding up misconception assessment research and curriculum design research.**

***Index Terms*—Data Science, Misconception Categorization, Assessment and Evaluation, Learning Analytics, Infrastructure and Technologies for Engineering Education, Computer-Based Learning and Courseware Technologies, Engineering Education Research**

I. Introduction

The U.S. Department of Education has long established that students must meet the requirements associated with their current K-12 educational standards in order to be competent in their subsequent grades. According to Pellegrino, Chudowski, and Glaser [1] students tend to develop knowledge gaps due to misconceptions. These misconceptions are the result of incorrect delivery of instruction and assessment techniques. When instruction is not accompanied by cognitive assessment, these misconceptions are hard to detect [1].

Chi et. al. [2] [3] [4] [5] [6] describe how science phenomena can be ontologically categorized and learners acquire misconceptions when they incorrectly categorize concepts. Over the past two decades research by Chi et. al. [2] [3] [4] [5] [6] [7] have documented a systematic method for determining which ontological category a student has assigned to a concept to determine its correctness. This process is called the predicate test. Predicate tests consist of extracting student predication used to describe a concept and comparing it to correct predication used by experts. By examining the verbal predication found in student descriptions of answers to multiple-choice questions teachers can identify each student's mental categorization of concepts, which in turn complements targeted curriculum design [2] [3] [4] [5] [6] [7].

A major issue of predicate test research and use is the extensive manual labor involved in its execution and the lack of trained personnel that are qualified to perform a proper predicate test.

The use of text classification to automate predicate tests can significantly reduce the amount of manual labor a group of teachers has to perform in lack thereof. Additionally a machine learning model could be deployed as an integral part of a student model building tool in a modern Intelligent Tutoring System (ITS) that focus on curriculum design based on individual student mental models.

This paper details the creation of a machine learning model to be used in the automation of predicate tests for the purpose of reducing the time it takes educators to perform individualized assessment of students and design curriculum content accordingly.

During our research the predicate test was considered a text classification task. Existing machine learning algorithms were trained using the dataset gathered by Chi in [7] from a public institution in the Midwestern U.S.. It was collected as part

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

of a research project where the predicate test was manually performed by educational engineering experts on engineering students studying the topics of dynamics and heat transfer. The dataset consists of textual descriptions written by students to describe their selection of a multiple-choice answer. Words and phrases that were identified by experts as ontological attributes belonging to the *emergent* process category or the *sequential* process category were labeled accordingly.

Our work focused on answering the following experimental questions:

1) Can the selected classifiers learn to correctly assign textual descriptions to the *emergent* and *sequential* ontological categories proposed by Chi in [2]?
2) Which of our chosen word tokenizing algorithms are best suited to extract features of the *emergent* and *sequential* ontological categories proposed by Chi in [2]?
3) Considering Support Vector Machines, Multinomial Naïve Bayes, Logistic Regression, and Bayesian Logistic Regression, which machine learning algorithm best learns to classify learner descriptions about concepts into the *emergent* and *sequential* ontological categories?
4) Can a single model yield the best performance for both the *emergent* and *sequential* ontological categories or is there a need to ensemble models to optimize predictions for both categories?

To perform our experiments textual explanations of a concept and their corresponding expert labels were the only features selected from the raw dataset. We chose to use the *sequential* and *emergent* process categories as in [7].The feature space was created using transformation, extraction and selection algorithms provided by WEKA [8].

We trained two models with each of the following classifiers: Support Vector Machines, Multinomial Naïve Bayes, Logistic Regression, and Bayesian Logistic Regression. Eight models were built — two for each classifier using Java's standard WordTokenizer and WEKA's N-Gram Tokenizer respectively. Our models were evaluated using 10-fold cross-validation. It is beyond the scope of this paper to discuss these classifiers in detail.

The selection of our best performing classification model was based on the resulting measures of accuracy, F-measure, and kappa coefficient.

Our results show that the predicate test proposed by Chi can be automated using text classification. The best performing model was trained using SVM with features transformed using an N-Gram Tokenizer with 3-grams and feature extraction by means of Term Frequency-Inverse Document Frequency.

Additional information about Chi's predicate test is found in the next section. An brief explanation follows describing what lead us to the consideration the predicate test a task of text classification, then our methodology and results follow. Implications and future work are discussed last.

II. Conceptual Change in Science Education

Science education research focuses on how learners acquire knowledge about science and how that knowledge is applied. Within the domain of conceptual change research two broad perspectives have emerged to describe the nature of knowledge structure coherence, misconceptions, and conceptual change. These are known as the *knowledge-as-theory* perspectives and the *knowledge-as-elements* perspectives [9]. These perspectives state that science domain knowledge acquired by learners can be broadly described, respectively, as unified frameworks with coherent theoretical structure or as independent collections of elements [9].

This research is aligned with the *knowledge-as-theory* perspectives, specifically the perspective documented by Chi et. al. in [2] [3] [4] [5] [6] [7]. Chi et. al.'s knowledge structure theory describes how science concepts can be ontologically categorized by learners and how incorrect categorization of concepts gives way to misconceptions and the need for conceptual change.

This remainder of this section describes the role of ontologies in misconceptions and provides a formal definition of conceptual change theory. It also includes a discussion about predicate tests and their use in conceptual change assessment.

A. Ontologies and Conceptual Change in Science Education

Conceptual change in the past two decades has become aligned with the notion that novice and expert understanding of concepts is based on ontological categories. When a learner incorrectly categorises a concept the he or she is said to have a misconception of that concept. This theory has been prominently documented by Chi, Slotta, et. al. in [2] [3] [4] [5] [6] [7].

The definition of conceptual change established by Chi, Slotta, et.al. is based on a combination of accepted positions within conceptual change literature and is an attempt to systematically diagnose and assess conceptual change. Three types of conceptual change have been defined by Chi: belief revision, mental model transformation, and categorical shift [6]. These three types of conceptual change are all based on the assumptions that entities in the world can be ontologically categorized, that the nature of physics science concepts dictates their categorization into *constraint-based interaction* concepts, and that students hold naïve preconceptions aligned with *substance-based* descriptions of concepts.

A contradicting theory documeted by Gupta et. al. in recent years [10] [11] [12] [13] has surfaced which proposes that associating physics science concepts to a single ontological category can be detrimental in the development of expertise. Gupta et. al. claim that novices do cross ontological boundaries when describing physics science concepts. Also they claim that resources which have been associated with a specific ontological category can be utilized to help teach concepts from different ontological categories as well [12]. Gupta et. al. argue that students do not have rigid ontological commitments, rather they can switch from matter based understanding to process based understanding of concepts, although most novice categorization of concepts tends to be substance based [12]. Gupta et. al.'s evidence is based on the inspection of expert literature and novice description of concepts. Although their findings are

theoretically coherent, their research does not explain how it is acceptable for experts to use *substance-based* descriptions for *constraint-based interaction* concepts. In other words, the fact that expert literature contains *substance-based* conceptions or examples of *constraint-based interaction* concepts proves that metaphors can be used to describe *constraint-based interaction* concepts with *substance-based* descriptions. In addition, this does not prove that concepts which belong to the *constraint-based interaction* category can also belong to the *substance* category [3]. For these reasons we subscribe to Chi's definition of conceptual change.

The following section of this document is dedicated describing the role of ontologies in Chi's conceptual change theory. Following that section it is noted why some types of conceptual change are apparently difficult. The next section details how the predicate test can be used to determine if and when conceptual change must take place.

1) Three Suppositions of Conceptual Change: As stated before, this work focuses on describing Chi's theory of conceptual change, which is based on three assumptions. The first is an epistemological assumption about the natural categorization of entities in the world. The second is a metaphysical assumption that describes how most science concepts belong to the process or constraint-based interactions category. The third is a psychological assumption related student's naïve preconceptions [3]. This section provides a brief description of those three assumptions.

a) Epistemological Proposition of Conceptual Change: The epistemological proposition of Chi's theory of conceptual change states that entities in the world naturally belong to certain major ontological categories, these being matter, processes, and mental states [3] (refer to Figure 1 for an example of this categorization of entities). The theory is loosely aligned with the exact names of those three major categories and states that more than three can exist, but focuses on the matter and processes categories or their equivalent descriptions [6].

Ontological attributes are those properties that an ontological category can possess due to being associated to that category. The following example aims to explain this notion: considering a shoe as an artifact from the matter category it can be said that the shoe must have defining attributes like a sole, most frequently has characteristic attributes like laces, and can potentially be worn, which is an ontological attribute [3].

Ontological trees are considered distinct if they possess mutually exclusive ontological attributes [3] [6]. In other words, ontological attributes from a given ontological tree cannot be applied to categories that belong to a different tree. For example, entities categorized as matter have ontological attributes such as"storable" and "having color, volume, or mass", which are attributes that cannot be assigned to a process category. In a similar way, ontological attributes such as "resulting in" or "occurring over time" can only be assigned to entities that form part of the processes tree. Categories within the same ontological tree can also be ontologically distinct when their respective categories cannot be shared between them [3] [6].

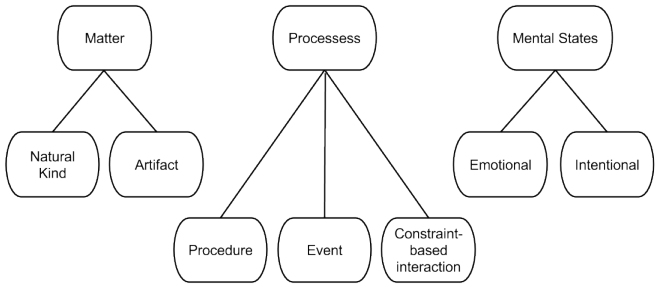

Fig. 1. Three major ontological trees from Chi et. al.'s epistemological proposition of conceptual change theory. Adapted from [3].

Ontological categorization is the basis of Chi's theory of conceptual change since conceptual change happens when students shift their categorization of concepts from one distinct ontological category to another [3]. This category shift by students can occur across major ontological trees or within them [3] [6].

b) Metaphysical Proposition of Conceptual Change: The second proposition of Chi's conceptual change theory is about the nature of science concepts. It exposes how the constraint-based interaction category, a subcategory of the processes ontological tree, can be used to describe many science concepts [3] [6]. Constraint based interactions are unpredictable and emergent with no definite beginning or end. This category encompasses concepts such as heat, electric current, and light, which are processes that cannot be assigned to any subcategories of the matter ontological tree [3] [6]. For example, electric current exists when a charged particle moves through an electric field. Thus the electric current process is considered a constraint-based interaction which emerges from the interaction of various components that belong to the matter ontological tree such as particles, wires, and batteries [3]. Since constraint-based interactions involve components from the matter category this can be confusing for students [3]. This assumption can apply to concepts outside of the physical sciences [3].

c) Psychological Proposition of Conceptual Change: The third proposition of Chi's conceptual change theory states that students hold naïve knowledge or preconceptions about science concepts [3] [6]. It explains the nature of some science misconceptions through a psychological point of view. In short, students' naïve conceptions of physical science concepts tend to consist of assigning process concepts to the matter category.

These preconceptions can exist at the proposition and mental model levels. Naïve knowledge at the proposition level is simple to remove, is referred to as non-robust misconceptions, and is corrected with conceptual reorganization. In other words, those preconceptions can be easier to correct or remove [3] [6]. Some naïve knowledge can be resistant to creative types of pedagogy approaches and is referred to as robust misconceptions [3] [6]. Diagnosing the presence or absence

of preconceptions in individual students reveals information about the incorrect category to which those preconceptions have been assigned by students.

B. Predicate Tests and Conceptual Change Assessment

Contributions by Chi and her colleagues Slotta, deLeuw, Santiago, et.al. have documented the use of the predicate test as the means to assess conceptual change, i.e. concept re-categorization at the ontological tree level [2] [5] [7]. The predicate test theory states that verbs used by students and experts to describe concepts correspond to ontological attributes of those concepts. Decades of research document how novices use *matter-based* predicates to describe *substance-based* concepts and *constraint-based interaction* concepts with the same frequency [2] [3] [4] [5] [6] [7]. Since it is usual for novice predication to contain ontological attributes of the wrong category predicate tests consist of analyzing verbal predicates used by students and contrasting them to the predicate use of experts.

It is also noted that expert predicate use contains a high frequency of *substance-based* predicates when describing *substance-based* concepts. In addition, the *predicate use profiles* reveal high frequency of *constraint-based interaction* predicate use for *constraint-based interaction* concept descriptions [2] [5] [7]. This is aligned with psychological proposition of conceptual change, which considers that novices hold naïve preconceptions causing the alignment of their mental models with *substance-based* predication for *constraint-based interaction* physics concepts [3]. In contrast, the proposition also states that experts consistently use *constraint-based interaction* predicates for *constraint-based interaction* descriptions of physics concepts.

Slotta et. al. developed expert taxonomies based on expert explanations of *substances* and *processes*. The taxonomies describe predicates in the form of single or multiple word phrases or ideas that are explicitly associated to certain ontological attributes and the ontological categories to which the associated ontological attributes belong to [2]. These taxonomies are used as the main criteria for predicate tests. Refer to Table II-B for a snippet of the substance and process predicate taxonomies.

TABLE I
SNIPPET OF SUBSTANCE AND CONSTRAINT-BASED INTERACTION PREDICATE TAXONOMIES PROPOSED BY SLOTTA ET. AL. IN [2].

Substance Predicates	Process Predicates
block	movement process
move	excitation
consume	equilibrium seeking
quantify	systemwide
accumulate	simultaneous
equivalent amounts	transfer

The following describes the events that lead up to and occur after predicate tests take place. Given a carefully crafted multiple-choice question about a certain science topic (e.g., dynamics, heat transfer) the students are required to explain their selected answer. The textual description of their answer is then examined to isolate predicates and the ontological attributes they contain. Then the isolated words, phrases, and sentences are classified as belonging to a certain ontological category that may or may not be the category to which the concept being described belongs to. This last step is considered to be the actual predicate test. In other words the predicate test does not include crafting multiple-choice questions whose answers can best reveal mental models of whoever answers them, administering the questions, or isolating predicates from textual descriptions the selected multiple-choice question answer. What it does include is the analysis of determining which of Chi's ontological categories for science concepts the student's predication corresponds to.

The resulting predicate test analysis is used to demonstrate the robustness of student's incorrect ontological categorizations by determining whether conceptual change should take place at the ontological tree level and to what degree [2] [5] [7].

III. RELATED WORK

Misconception assessment using Machine Learning techniques has found its place in the construction of student models within Intelligent Tutoring Systems (ITS). In addition student models play a major role in ITS' personalization strategies. This is due to the fact that ITS incorporate adaptive learning techniques based on each individual student's knowledge [14]

This work focuses on student misconception detection at the ontological level to develop student models that describe their current ontological categorization of concepts. In other words we have focused on categorizing misconceptions. We are not aware of any work that uses our machine learning approach to model erroneous conceptions in students.

The the rest of this section mentions research work that has a close resemblance to our work, although they do not address the problem of categorizing misconceptions. Instead they mainly focus on the presence of misconceptions in students or describing the general knowledge possessed by students. In addition, they focus on determining which learning materials are needed to correct student misconceptions or to advance to subsequent lessons respectively.

In the research documented by Liu in [15] the proposed system is already aware of the possible misconceptions that can arise while learning statistics topics. The system focuses on evaluating students to determine if the possess any of these previously determined misconceptions and providing feedback to make them aware of their existence.

In order to advance to subsequent lessons Wang used a pre-test and a two-tier tests to identify knowledge gaps and dynamically select additional learning materials a student may need [16].

Ehimwenma proposed the use of a multi-agent system to determine which concepts have not been learned by students in [17]. Pre-assessment strategies are employed to model current knowledge in students, then knowledge gains and gaps are identified to determine which learning materials are recommended for learners to able to advance to further lessons.

Bayesian Student Models (BSMs) have also been used by Millán in [18] to develop student models. They are based on using knowledge and evidence variables, as well as their correlation, to construct Bayesian Networks. In other words the network structure or nodes are elements that represent whether or not students have knowledge about a specific domain, and the answer to questions about that domain. In addition the edges between these two types of nodes is used to describe their correlation for a particular student.

Wenando in [19] and Rus in [20] both used student-written paragraphs in their construction of mental models. Their work is similar to our research due to their use of text written by students. However the mental models that can be inferred by the output of their classification consist of approximations of how commensurate student essays are with regards to expert-written essays, taxonomies, and predication.

We have observed that while the previously mentioned research can model student knowledge and identify knowledge gaps, they are not focused on misconception categorization. The main difference between our work and the research performed by [19] and [20] lies in how we train a model to learn which category a learner has assigned to a concept instead of learning to output a measure of incommensurability between student text and text written by experts.

IV. Text Classification of Student Predicate Use

Consider a student's textual explanations of multiple choice science questions as the set of documents D to be categorized into the set of ontological categories C used by Chi. According to text classification theory [8] [21] [22] [23] the mentioned documents could be classified into each category using text classification.

The classification to be performed consists of assigning student explanations about science topics to one or a combination of the major categories of concepts involved in learning science. This is the approach taken by Chi in [2] [3] [4] [5] [6] [7]. Their research has identified the three main categories used in science concept discussions as [6] substances, sequential processes, and emergent processes. These categories can be used as our text classification labels.

V. Methodology

The Waikato Environment for Knowledge Analysis (WEKA) toolkit [8] main tool we used to accomplish our text classification tasks. Our training set was prepared using the data collected by Chi et. al. in [7], which consists of textual explanations to multiple-choice science questions administered as pre- and post-questionnaires. The data was collected as part of a research project where the predicate test was manually performed by educational engineering experts. Predicate test results from the pre-questionnaire were used to construct mental models of students and design curriculum content accordingly. Among the results of that research is an expert annotated dataset where the verbs or phrases used by students in their explanations are identified as ontological attributes. Furthermore those ontological attributes were categorized into the *emergent process*, *direct process*, and *emergent and sequential process* or *mixed* categories by educational engineering experts as well.

The student sentences that are being considered for classifier training have not been tampered with by experts. This suggests that a classifier can be trained with expert labeled predicate test data and compute predicate test results on student textual descriptions of science concepts that have no expert labels. Using that approach to compute predicate test results could be considered a fully automated predicate test.

The bullet list found below shows textual descriptions written by students to describe their selected answer to multiple-choice questions about diffusion and heat transfer. It was extracted from the original dataset under consideration. According to Chi's predicate test theory each sentence contains *sequential* and *emergent* phrases, which reveal information of each student's mental categorizations of concepts. These phrases were identified by experts and bolded if they were *sequential* phrases and bolded and italicized if they were determined to be *emergent* in nature.

- Stirring the water **causes** more molecular motion and allows more salt **become evenly dispersed throughout** the container
- The air cannot escape and helium can (as stated above), therefore due to ***random movement*** of helium, some helium is likely to escape in the process
- ***The rates will be the same*** because the non stirred glass **will eventually reach equilibrium** through diffusion and have the same temperature as the stirred glass.
- The ***random*** motion of the dye molecules **causes them to collide** and **move into the beaker** with just the water
- Thermal excitation makes molecules move faster, therefore there is an increase of the ***molecules random motion***, therefore the concentration of the dye **reaches equilibrium** with the water quicker in a heated beaker

Since our available dataset showed the presence of more than one conceptual schema used by students to describe concepts we chose to have three categories in our dataset: *emergent*, *sequential*, and *mixed*. In our dataset the samples that belong to both the *emergent* and *sequential* classes were labeled as *mixed*. Our experiment performs multi-class classification of instances into those three categories.

Our preprocessing step was manually completed by constructing a file in ARFF format, which is the document format used by WEKA. Each labeled sentence was extracted from the original dataset to yield a total of 40 instances of the *emergent* class, 99 instances of the *sequential* class, and 50 instances of the *mixed*.

The first WEKA filter applied to our dataset for feature transformation is known as StringToWordVector. This filter

tokenizes each string attribute in our data samples into a set of features consisting of each word in the string and information about word occurrence [24] [25]. We used Java's standard WordTokenizer and WEKA's N-Gram tokenizer. In addition the StringToWordVector was flagged to perform feature extraction considering Term Frequency-Inverse Document Frequency (TF-IDF).

The second WEKA filter applied to our dataset is known as the Synthetic Minority Over-Sampling Technique (SMOTE) [26]. This algorithm performs feature selection by considering minority class instances and their *k* nearest neighbors to generates synthetic data. Synthetic samples are generated in three steps .The first calculates the difference between a minority class feature vector and its chosen nearest neighbor. Then the difference is multiplied by a random number between 0 and 1 and the result is added to the feature vector under consideration [26].

The list below describes all the parameters that were changed from default when applying StringToWord vector during our experiments:

- IDFTransform - set to true in order to transform word frequencies in a document into:

$$TF\text{-}IDF(t_i, d_j, D) = TF(t_i, d_j)\dot{I}DF(t_i, D); \quad i = 1, ..., m, j = 1, ..., n \tag{1}$$

  where $TF(t_i, d_j)$ is the number of times term t_i appears in document d_j, D is the data corpus under consideration, and

$$IDF(t_i, D) = \log \frac{N}{|\{d_j \in D : t_i \in d_j\}|} \tag{2}$$

  where $N = |D|$ (total number of documents) and $\{d_j \in D : t_i \in d_j\}$ = number of documents where term t_i is present
- normalizeDocLength - set to true to have the word frequencies for each document normalized
- tokenizer - selects the tokenizing algorithm to use on the strings
- wordsToKeep - set to keep 10,000 to words per class

We used SMOTE to correct our class imbalance. *Emergent* class examples were randomly oversampled by 150% with 5 nearest neighbors to obtain a total of 100 instances from the *emergent process* class. *Mixed* class examples were randomly oversampled by 100% with 5 nearest neighbors to obtain a total of 100 instances from the *mixed* class.

To answer our experimental questions eight text classification experiments were performed using the following machine learning algorithms with their respective default parameters in WEKA:

- Support Vector Machines (SVM)
- Multinomial Naïve Bayes (MNB)
- Logistic Regression (LR)
- Bayesian Logistic Regression (BLR)

The first four experiments consisted in using Java's WordTokenizer for feature transformation with and without TF-IDF in our feature extraction step. The last four experiments consisted of using WEKA's N-Gram tokenizer for feature transformation with bags of up to three words. All algorithms were trained to learn two target classes — *emergent* and *sequential*.

We used 10-fold cross-validation to evaluate the aforementioned experiments. We also chose to describe relevant results in terms of accuracy and kappa coefficient (κ).

VI. Experiment Results

Table II shows the measured accuracy of each constructed model, table III shows each model's F-measure score, and table IV shows each model's kappa coefficient. Tables V, VI, VII, VIII show confusion matrices for each classifier's best performance.

TABLE II
ACCURACY PER TOKENIZER FOR EACH CLASSIFIER USING ON 10-FOLD C-V

Tokenizer	SVM	MNB	LR	BLR
Java	90.04%	91.54%	88.05%	91.54%
3-gram	97.51%	93.53%	93.53%	97.01%

TABLE III
F-MEASURE PER CLASS AND TOKENIZER FOR EACH CLASSIFIER USING ON 10-FOLD C-V

Tokenizer	Class	SVM	MNB	LR	BLR
Java	Emergent	0.91	0.92	0.89	0.92
Java	Sequential	0.89	0.91	0.87	0.91
3-gram	Emergent	0.98	0.94	0.94	0.97
3-gram	Sequential	0.97	0.93	0.93	0.97

TABLE IV
KAPPA COEFFICIENT PER TOKENIZER FOR EACH CLASSIFIER USING 10-FOLD C-V

Tokenizer	SVM	MNB	LR	BLR
Java	0.80	0.83	0.76	0.83
3-gram	0.95	0.87	0.87	0.94

TABLE V
CONFUSION MATRIX FOR SVM USING 3-GRAM TOKENIZER AND 10-FOLD C-V

	Emergent	Sequential
Emergent	102	0
Sequential	5	94

VII. Results Discussion

The scope of this research focuses on an approach to achieve predicate test automation with text classification techniques. According to the results presented in Section VI the proposed

TABLE VI
CONFUSION MATRIX FOR MNB USING 3-GRAM TOKENIZER AND 10-FOLD C-V

	Emergent	Sequential
Emergent	101	1
Sequential	12	87

TABLE VII
CONFUSION MATRIX FOR LR USING 3-GRAM TOKENIZER AND 10-FOLD C-V

	Emergent	Sequential
Emergent	101	1
Sequential	12	87

TABLE VIII
CONFUSION MATRIX FOR BLR USING 3-GRAM TOKENIZER AND 10-FOLD C-V

	Emergent	Sequential
Emergent	102	0
Sequential	6	93

system can successfully learn to categorize sentences into the *emergent* and *sequential* classes.

When features were extracted using Java's standard WordTokenizer Multinomial Naïve Bayes and Bayesian Logistic Regression models outperformed models trained using SVM and Logistic Regression. Both of these models scored an accuracy of 91.54%, F-measure score of 0.92 and 0.91 for the *emergent* and *sequential* classes respectively, and kappa of 0.83.

As a result of applying the 3-gram tokenizing technique we observe an increase in all marks of performance for each classifier. This observation reveals how the 3-gram tokenizer is best suited for features extraction of the *emergent* and *sequential* ontological categories proposed by Chi in [2] in comparison to Java's tokenizer.

Amongst the results of using the 3-gram tokenizer we observe how our SVM del outperformed all other models with an accuracy of 97.51% and F-measure of 0.98 and 0.97 for the *emergent* and *sequential* classes respectively. SVM combined with a 3-gram tokenizer also resulted in a kappa score of 0.95, which is considered a very high kappa score according to [27] [28] [29]. The SVM model also inferred the highest amount of correctly categorized instances for both the *emergent* and *sequential* classes — there would be no need ensemble models to obtain the best performance for both the *emergent* and *sequential* ontological categories.

In addition we have also observed that the *emergent* ontological category obtained the least amount of false negatives. This might be because features from this category have reduced variance in our small datasets since *emergent* phrases are usually very similar.

VIII. CONCLUSIONS

This research is based on building text classification models for misconception assessment. We subscribe to Chi et.al's misconception theory, which arguments that learners assign newly acquired knowledge about a science concept to an ontological category that best describes the nature of the such concept. Misconceptions occur when learners assign an incorrect category to concepts under study. In addition Chi has developed a process known as the predicate test, which aids in the assessment of misconceptions by inspecting student textual descriptions of science concepts and categorizing the descriptions as belonging to an *emergent* process, a *sequential* process, or both. This process is known to be a time consuming task that can only be performed by trained experts from the educational engineering domain.

The main purpose of this research was to show how predicate tests could be automated with text classification models using a previously annotated dataset. This approach reduces the time it takes to deliver predicate test results, which are used by researchers and instructors in the development of customized curriculum content. Chi's predicate test was automated by applying the text classification to data acquired from a pre-course questionnaire designed to gather data about a student's mental categorization of concepts. Four different classifiers were trained in order to show that the *emergent* and *sequential* process categories can be successfully learned.

The dataset was collected from engineering students enrolled at a public institution in the Midwestern U.S.. Student explanations to multiple choice question answers were considered documents for classification belonging to two categories: *sequential* and *emergent*. Furthermore Java's WordTokenizer and WEKA's N-Gram tokenizer were used with 3-grams for feature transformation and TF-IDF for feature extraction to determine which is best suited for predicting those ontological categories. The classification models were trained using WEKA's Support Vector Machines, Multinomial Naïve Bayes, Logistic Regression, and Bayesian Logistic Regression implementations and evaluated with measures of accuracy, F-measure, and kappa coefficients.

According to our classification results we have concluded that it is possible to build a classification model of acceptable performance using complete student descriptions about concepts as the only source of input for our feature space creation. We also found it is possible to classify learner descriptions about concepts into the *emergent* and *sequential* ontological categories proposed by Chi [7]. We have also observed improved performance with all classifiers using WEKA's N-Gram tokenizer with 3-grams to vectorize words. Furthermore, we have determined that our best performing classification model was trained with SVM using data tokenized with 3-grams. Our best performing SVM model scored an accuracy of 97.51%, F-measure of 0.98 and 0.97 for the *emergent* and *sequential* classes respectively, and a kappa coefficient of 0.95. This SVM

model also resulted in the least amount of incorrectly classified instances when compared to the other three classifiers.

A positive contribution to the science of educational engineering has resulted from the combination of Chi's predicate tests with text classification. With the intent to automatize Chi's predicate tests we have documented a methodology for exploring the possibility of helping educators design customized curriculum content using real-time, individualized misconception assessment.

ACKNOWLEDGMENT

Presentation of this work has been supported in part by NSF Grant CNS-1042341.

We thank collaborators who provided the dataset used in our research, including but not limited to Dazhi Yang, Ruth Streveler, James Slottta, and Ronald Miller. Their work was supported by the National Science Foundation as part of a project named "Developing Ontological Schema Training Methods to Help Students Develop Scientifically Accurate Mental Models of Engineering Concepts" under grant no. EEC-0550169.

REFERENCES

[1] James W Pellegrino, Naomi Chudowsky, Robert Glaser, et al. *Knowing what students know: The science and design of educational assessment*. National Academies Press, 2001.

[2] James D Slotta, Michelene TH Chi, and Elana Joram. Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change. *Cognition and instruction*, 13(3):373–400, 1995.

[3] Michelene TH Chi, James D Slotta, and Nicholas De Leeuw. From things to processes: A theory of conceptual change for learning science concepts. *Learning and instruction*, 4(1):27–43, 1994.

[4] Michelene TH Chi and Rod D Roscoe. The processes and challenges of conceptual change. In *Reconsidering conceptual change: Issues in theory and practice*, pages 3–27. Springer, 2002.

[5] James D Slotta and MT Chi. How physics novices can overcome robust misconceptions through ontology training. *Manuscript submitted for publication*, 1999.

[6] Michelene TH Chi. Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. *International handbook of research on conceptual change*, pages 61–82, 2008.

[7] Dazhi Yang, Aidsa Santiago Roman, Ruth A Streveler, Ronald L Miller, James Slotta, and Michelene Chi. Repairing student misconceptions using ontology training: A study with junior and senior undergraduate engineering students. In *Proceedings of the 2010 ASEE Annual Conference and Expo*, June 2010.

[8] Abdullah H Wahbeh and Mohammed Al-Kabi. Comparative assessment of the performance of three weka text classifiers applied to arabic text. 2012.

[9] Gökhan Özdemir and Douglas Burton Clark. An overview of conceptual change theories. *Eurasia Journal of Mathematics, Science & Technology Education*, 3(4):351–361, 2007.

[10] Ayush Gupta and Andrew Elby. Beyond epistemological deficits: dynamic explanations of engineering students difficulties with mathematical sense-making. *International Journal of Science Education*, 33(18):2463–2488, 2011.

[11] Ayush Gupta, Andrew Elby, and Luke D Conlin. How substance-based ontologies for gravity can be productive: A case study. *arXiv preprint arXiv:1305.1225*, 2013.

[12] Ayush Gupta, David Hammer, and Edward F Redish. Towards a dynamic model of learners' ontologies in physics. In *Proceedings of the 8th international conference on International conference for the learning sciences-Volume 1*, pages 313–318. International Society of the Learning Sciences, 2008.

[13] Ayush Gupta, David Hammer, and Edward F Redish. The case for dynamic models of learners' ontologies in physics. *the journal of the learning sciences*, 19(3):285–321, 2010.

[14] Vladan Devedzic. Education and the semantic web. *International Journal of Artificial Intelligence in Education*, 14(2):165–191, 2004.

[15] Liu Tzu-Chien. Developing simulation-based computer assisted learning to correct students' statistical misconceptions based on cognitive conflict theory, using "correlation" as an example. *Journal of Educational Technology Society*, 13(2):180 – 192, 2010.

[16] Tzu-Hua Wang. Developing an assessment-centered e-learning system for improving student learning effectiveness. *Computers Education*, 73:189 – 203, 2014.

[17] Kennedy E Ehimwenma, Martin Beer, and Paul Crowther. Adaptive multiagent system for learning gap identification through semantic communication and classified rules learning. In *7th International Conference on Computer Supported Education. In Doctoral Consortium (CSEDU)*, pages 33–38, 2015.

[18] Eva Millán, Guiomar Jiménez, María-Victoria Belmonte, and José-Luis Pérez-de-la Cruz. Learning bayesian networks for student modeling. In *Artificial Intelligence in Education*, pages 718–721. Springer, 2015.

[19] Vasile Rus, Mihai Lintean, and Azevedo Roger. Automatic detection of student mental models during prior knowledge activation in metatutor. pages 161–170, 01 2009.

[20] Febby Apri Wenando, Teguh Adji, and Igi Ardiyanto. Text classification to detect student level of understanding in prior knowledge activation process. 23:2285–2287, 03 2017.

[21] Duan Li-guo, Di Peng, and Li Ai-ping. A new naive bayes text classification algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 12(2):947–952, 2014.

[22] Frans Coenen, Paul Leng, Robert Sanderson, and Yanbo J Wang. Statistical identification of key phrases for text classification. In *Machine Learning and Data Mining in Pattern Recognition*, pages 838–853. Springer, 2007.

[23] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[24] M Jayakameswaraiah and S Ramakrishna. Development of data mining system to analyze cars using tknn clustering algorithm. *International Journal of Advanced Research in Computer Engineering Technology*, 3(7), 2014.

[25] V Umadevi. Sentiment analysis using weka. *International Journal of Advanced Research in Computer Engineering Technology*, 18(4), 2014.

[26] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

[27] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[28] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973.

[29] Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990.