QUESTION CLASSIFICATION BASED ON BLOOM'S TAXONOMY COGNITIVE DOMAIN USING MODIFIED TF-IDF AND WORD2VEC

MANAL MOHAMMED AL-TAMIMI

UNIVERSITI KEBANGSAAN MALAYSIA

QUESTION CLASSIFICATION BASED ON BLOOM'S TAXONOMY COGNITIVE DOMAIN USING MODIFIED TF-IDF AND WORD2VEC

MANAL MOHAMMED AL-TAMIMI

DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

KLASIFIKASI SOALAN BERDASARKAN DOMAIN KOGNITIF TAKSONOMI BLOOM MENGGUNAKAN PENGUBAHSUAIAN TF-IDF DAN WORD2VEC

MANAL MOHAMMED AL-TAMIMI

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH
IJAZAH SARJANA SAINS KOMPUTER

FAKULTI TEKNOLOGI DAN SIANS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

13 August 2018

MANAL MOHAMMED AL-TAMIMI
P82838

ACKNOWLEDGEMENT

First and foremost, praise be to Almighty Allah for all his blessings for giving me patience and good health throughout the duration of this master research.

I am grateful to my supervisor Prof. Dr. Nazlia Omar, for her guidance, patience, valuable suggestions and supervision

Special thanks to my parents, my sister, and my brothers, for their love, supporting, and for their prayers…

I am thankful to my beloved friend Ashwaq Qasem for her ultimate support.

My sincere thanks to all postgraduate staff and students of FTSM for their help, friendship, and creating a pleasant working environment throughout my years in UKM.

I would also like to acknowledge my sponsors, Hadhramout Establishment for Human Development, for granting me this outstanding opportunity to obtain my Master's degree, and for their financial support.

Finally, thanks to everybody who contributed to this achievement either directly or indirectly.

ABSTRACT

Examination question assessment plays an important role in educational institutes, since it is one of the most common method to evaluate student's achievement in specific course. Therefore, there is a crucial need to write a balanced and high-quality exam, which satisfy different levels of cognitive. Thus, many lecturers use Bloom's taxonomy cognitive domain, which is a popular framework developed for the purpose of assesses students' intellectual abilities and skills. However, the process of classifying questions automatically based on Bloom's taxonomy is a challenging task due to the shortness of questions. Therefore, several works have been done to automatically classifying questions in accordance to Bloom's taxonomy. Most of these works classify questions in a specific domain, where there is a lack of techniques on classifying question over multi-domain area. The aim of this study is to build a generic question classification model to classify question based on Bloom's taxonomy cognitive domain from several areas. This study proposed a new method for classifying questions automatically by extracting two features, namely TFPOS-IDF and pre-trained word2vec. The purpose of first feature is to calculate the term frequency- inverse documents frequency based on part of speech, in order to assign a suitable weight for important words in the question. While pre-trained word2vec, the semantic feature, used to boost and enhance the classification process. Then, the combination of these both features are fed into Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classifiers, in order to classify the questions. The experiments have used two dataset. The first dataset contains 141 questions, while the other dataset contains 600 questions. The questions in both dataset are collected from different domains, and divided into 80% training set and 20% test set. The classification result for the first dataset achieves an average of 83.7% and 71.1% weighted F1-measure respectively. While the classification result for the second dataset achieves an average of 89.7% and 85.4% weighted F1-measure respectively. The finding from this study showed that the proposed method is significant in classifying questions from multiple domain.

ABSTRAK

Penilaian soalan peperiksaan memainkan peranan penting dalam institusi pendidikan, kerana ia adalah salah satu kaedah yang paling biasa untuk menilai pencapaian pelajar dalam kursus tertentu. Jadi, terdapat keperluan penting dalam menulis soalan peperiksaan yang seimbang dan berkualiti tinggi, yang boleh menepati tahap kognitif yang berbeza. Oleh itu, ramai pensyarah telah menggunakan domain kognitif Taksonomi Bloom, yang merupakan rangka kerja popular untuk tujuan menilai kebolehan dan kemahiran intelektual pelajar. Walau bagaimanapun, proses mengklasifikasikan soalan secara automatik berdasarkan Taksonomi Bloom menjadi tugas yang mencabar jika kekurangan soalan. Oleh itu, beberapa kajian telah dilakukan untuk mengklasifikasikan soalan secara automatik mengikut Taksonomi Bloom. Kebanyakan ujikaji ini mengklasifikasikan soalan dalam domain tertentu, di mana terdapat kekurangan pada teknik dalam mengklasifikasikan soalan multi-domain. Tujuan kajian ini adalah untuk membina model klasifikasi soalan generik untuk mengklasifikasikan soalan berdasarkan domain Taksonomi Bloom dari beberapa bidang. Kajian ini mencadangkan kaedah baru untuk mengklasifikasikan soalan secara automatik dengan mengekstrak dua ciri, iaitu TFPOS-IDF dan word2vec pra-terlatih. Tujuan ciri pertama adalah untuk mengira frekuensi kekerapan istilah berdasarkan golongan kata, untuk memberi pemberat yang bersesuaian bagi perkataan yang penting di dalam soalan. Manakala ciri semantik word2vec pra-terlatih, digunakan untuk meningkatkan dan menambahbaik proses klasifikasi. Kemudian, gabungan kedua-dua ciri ini dimasukkan ke dalam teknik klasifikasi Mesin Vektor Sokongan (SVM) dan k-jiran terdekat (KNN), untuk mengklasifikasikan soalan. Eksperimen ini menggunakan dua set data. Set data pertama mengandungi 141 soalan, manakala set data yang kedua mengandungi 600 soalan. Soalan bagi kedua-dua set data dikumpulkan dari domain yang berlainan dan dibahagi kepada 80% set latihan dan 20% set ujian. Hasil klasifikasi untuk set data pertama mencapai purata masing-masing sebanyak 83.7% dan 71.1 % bagi ukuran F1 berpemberat. Manakala hasil klasifikasi untuk set data kedua pula mencapai purata masing-masing sebanyak 89.7% dan 85.4% bagi ukuran F1 berpemberat. Hasil kajian ini menunjukkan bahawa kaedah yang dicadangkan adalah berkesan dalam mengklasifikasi soalan dari pelbagai domain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CBOW	Continuous Bag-of-Word
CF-DF	Category Frequency-Document Frequency
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DF	Document Frequency
FP	False Positive
FN	False Negative
IDF	Inverse Document Frequency
KNN	K-Nearest Neighbours
MOOC	Massive Open Online Courses
NB	Naïve Bayes
NLP	Natural Language Processing
POS	Part-of-Speech
RBF	Radial Basis Function
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency- Inverse Document Frequency
TFPOS-IDF	Term Frequency Part Of Speech-Inverse Document Frequency
TP	True Positive
W2V-TFPOSIDF	Word2Vec- Term Frequency Part Of Speech InverseDocument Frequency

CHAPTER I

INTRODUCTION

1.1 RESEARCH BACKGROUND

Examination assessment plays an important role in evaluating how students' proficient in course content (Swart 2010). Writing high quality and balanced exam that satisfy different levels of cognitive is not an easy task (Haris & Omar 2015). That is why writing the examination in a comprehensive way taking into consideration the difficulty levels of questions, that matching the objectives and outcomes of the course in a standard way such as Bloom's taxonomy is a crucial task (Swart, 2010; Osman & Yahya, 2016).

Different assessment taxonomies are available such as Bloom's taxonomy and SOLO taxonomy (Jayakodi et al. 2016) that emphasis, guide and help instructors to evaluate students' achievement in the specific course. Bloom's taxonomy is very popular and widely acceptable because many academics are familiar with it. In addition, it can be applied to different kind of questions and various subjects (Abduljabbar 2015). Thus, it is a suitable framework for classifying examination questions.

Bloom's taxonomy involves three domains. Cognitive Domain, one of these three domains, which covers different thinking skills starting from simplest to the most complex one. Bloom's taxonomy cognitive domain consists of six levels: knowledge level, comprehension level, application level, analysis level, synthesis level and evaluation level. Knowledge level refers to the question that concerns about recalling and defining factual information. Comprehension level refers to the question that needs to be organized, compared and interpreted based on the understanding of the topics and previous knowledge. Application level refers to the question that concerns about solving

the new problem by using the gained knowledge. Analysis level refers to the question that requires the ability to determine and distinguish the relationship between different components. Synthesis level refers to the question that concerns about creativity and the combination of several ideas to produce a new solution. Evaluation level refers to the question that demonstrates the ability to defend and justify the quality of information according to a set of criteria (Kennedy 2006; Abduljabbar & Omar 2015).

Verb plays an important role in determining the level of the question in Bloom's taxonomy. Kennedy (2006) mentioned that, the key in writing learning outcomes by Bloom's taxonomy is verbs. Therefore, many universities such as the University of Central Florida[1] and Missouri State University[2] provides guidance documents for their academic staff in order to show them how to use these verbs to meet Bloom's taxonomy.

Recently, researchers (Yahya 2017; Jayakodi et al. 2016; Haris & Omar 2015; Kusuma et al. 2015) have shown an increased interest in automating evaluating examination based on Bloom's taxonomy cognitive domain. Several approaches have been used to achieve this goal, pure rule-based approach (Omar et al. 2012), machine learning techniques (Yahya & Osama 2011; Kusuma et al. 2015), and even the evolutionary algorithms (Yahya & Osman 2015; Yahya 2017) which are usually used to solve optimization problems. Many features have been extracted with these techniques such as lexical features, and syntactic features, while few of them used semantic features. On the other hand, most works are handled classifying questions from a specific domain, where there is a lack of techniques on classifying questions based on Bloom's taxonomy cognitive domain over the multi-domain area (Hussein 2017; Sangodiah et al. 2017).

Therefore, this study aims to build a generic question classification model based on Bloom's taxonomy cognitive domain. Hence, there still remains considerable room for further improvement, particularly in open domain area.

---

[1] http://www.fctl.ucf.edu/teachingandlearningresources/coursedesign/bloomstaxonomy/

[2] https://www.missouristate.edu/assets/fctl/Blooms_Taxonomy_Action_Verbs.pdf

1.2 PROBLEM STATEMENT

The importance of classifying questions regarding to Bloom's taxonomy cognitive levels lies in providing a suitable and appropriate way to measure students' intellectual abilities (Bloom 1956). Therefore, the automatic classifying of examination questions based on Bloom's taxonomy is most required, especially in an educational environment (Osman & Yahya 2016), since the process of classifying exam questions manually is time-consuming. Furthermore, some academicians have no idea about Bloom's taxonomy (Omar et al. 2012), or have no ability to distinguish the difference between Bloom's taxonomy's levels which may lead to misclassification. Hence, this may lead to poor quality examination (Omar et al. 2012 ; Jayakodi et al. 2016).

Various techniques have been used to tackle the problems of automatic classifying questions based on Bloom's taxonomy either by defining a set of rules (Omar set al. 2012; Haris & Omar 2015; Jayakodi et al. 2016), or by using machine learning techniques along with lexical, syntactic or semantic features (Yahya et al. 2013; Kusuma et al. 2015;Hussein 2017). Basically, the rule-based approach defines a set of rules to determine the appropriate level of question. The drawback of this approach is that it is time-consuming since many rules must be written manually to handle all cases, which is inefficient. Moreover, it is a static approach, which means it designed to fit and handle specific domain. Thus, it will not work for other domains. To overcome these problems researchers used machine learning since it is more dynamic and robust (Abduljabbar & Omar 2015).

In general, questions classification is unlike documents classification, since questions are written in short form. Documents classification aids users to get and extract useful information easily due to the extensive available information. Whereas, short text suffers from the lack of gained information and sparsity (Yang et al. 2013; Wang et al. 2016).

Thus, it is not suitable to use the pure statistical method in order to perform question classification such as N-gram and TF-IDF, since these techniques need a huge amount of data in order to get high accuracy (Abduljabbar & Omar 2015). In addition,

removing stop-words in preprocessing text in document classification is a common step to reduce insignificant words. Nevertheless, some of these stop-words such as *what, when, where*, and *how* are worthwhile in question classification process (Sangodiah et al. 2014).

Another main issue in classifying questions based on Bloom's cognitive is assigning the suitable weight for keywords that determine the level of the question, especially for the words that might appear in more than one taxonomy, such as Define that belong to *Knowledge level* and *Comprehension level*. To handle this issue, (Chang & Chung 2009; Omar et al. 2012) proposed assigning weight to the words from the perspective of experts. Using this method may lead to inconsistency, due to the variety of background knowledge of each expert. As a result, this caused poor performance of the classification process. On other hand, verbs are very important to distinguish the level of question. Using traditional weighting method such as original TF-IDF will also produce a poor outcome as in (Yahya & Osama 2011; Yahya et al. 2012).

Most studies in classifying question upon Bloom's taxonomy have been focused on a specific domain, while few studies have investigated under an open domain (Hussein 2017; Sangodiah et al. 2017). In order to tackle the problem of shortness questions length and handle a variety of domains, Hussein (2017) have used external semantic knowledge *WordNet* to expand questions. On other hand, still there are other methods which might not have been yet investigated regarding to classifying questions based on Bloom's cognitive level, such as using word embedding, e.g. word2vec which have shown good results in sentiment analysis and question classification in question answering system (Kim 2014; Dahou et al. 2016; Wang et al. 2016). Therefore, this study aims to take the benefit of using word embedding with a combination of improved statistical feature, to enhance the classification process.

1.3 RESEARCH OBJECTIVES

The aim of this study is to improve the process of classifying exams questions based on Bloom's Taxonomy under open domain. Therefore, the objectives of this research are stated as follows:

1. To develop a weighting method to set the priority of words based on modified TF-IDF with Part-of-Speech (POS).
2. To perform feature extraction based on pre-trained word2vec to enhance the classification process.
3. To evaluate the proposed method via machine learning classification algorithms.

1.4 RESEARCH SCOPE

The aim of this study is to build a generic question classification model based on Bloom's taxonomy cognitive domain. Accordingly, two open domain datasets are used. The first dataset is collected from several resources, while the second dataset is introduced by Yahya et al. (2012). Both of these datasets consist of open-ended questions. There are neither multiple choices nor true or false questions.

1.5 SIGNIFICANCE OF STUDY

Mainly written examination is one of the assessment techniques that used to determine the students achievement of learning outcomes (Kennedy 2006). Writing questions in a proper way, where different intellectual skills are taken into account is a challenging task. Therefore, using a framework such as Bloom's Taxonomy cognitive domain will lead to produce a suitable exam. The benefit of automating the process of classifying questions based on Bloom's taxonomy lies in saving academicians time and in using it effectively throughout different kind of applications such as automatic test generation systems, intelligent tutoring systems and even more in the serious game as (Rasim et al. 2016) work. However, this research proposes a model that improve statistical feature to assign suitable weights for the important words in question. In addition, to use semantic feature word2vec in order to enhance classification result.

1.6 DISSERTATION OVERVIEW

This research consists of five chapters, organized as follows:

Chapter II- Literature Review This Chapter explains in details the Bloom's taxonomy cognitive domain with illustrative examples. Furthermore, it includes a review of the used techniques in previous studies in classifying questions based on Bloom's Taxonomy and the feature extraction.

Chapter III- Research Method In this chapter the improved feature and the proposed method is explained in details. In addition, to the classification algorithms and metrics used to evaluate the proposed method.

Chapter IV- Experimental Results The aim of this chapter is to address the experimental settings and results which have been performed by two machine learning classifiers Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), with two datasets. Then to discuss and analyze the outcome of these experiments. After that, demonstrates the comparison between the performance of the classifiers, and the comparison against the related work. In addition, to check the significant test of the proposed method.

Chapter V- Conclusion and Future Work In this chapter the proposed study is concluded by providing an overall summary of what has been handled. Moreover, some suggestions will provide to extend this work in the future.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter contains the following sections: Section 2.2, defines Bloom's taxonomy cognitive level and provides examples for each level. Moreover, is briefly highlights the differences between the new and old version of Bloom's taxonomy cognitive domain. Section 2.3 discusses the different techniques used in question classification. Section 2.4 discusses the feature extraction in classifying questions and focuses more on the statistical feature TF-IDF. In addition, it provides a brief information about word embeddings, and more specifically about word2vec. Section 2.5 provides analysis about previous related works.

2.2 BLOOM'S TAXONOMY COGNITIVE DOMAIN

Benjamin Bloom and his team introduce Bloom's taxonomy in 1956 which basically involves three domains, the one that developed for the purpose of assesses students' intellectual abilities and skills known as Cognitive Domain (Bloom 1956). Bloom's taxonomy has gained a high attention, and widely applied in the educational field. In addition, it is translated into 22 different languages. Moreover, it is one of the most cited work in education (Forehand 2010).

Bloom's Taxonomy cognitive domain has a hierarchical structure including six levels namely knowledge level, comprehension level, application level, analysis level, synthesis level and evaluation level. According to Ghanem Nayef et al. (2013) knowledge level and comprehension levels considers as lower order thinking, whereas analysis level, synthesis level, and evaluation level categorize as higher order thinking.

Whilst application level lies in between, thus it belongs to both groups. However, swart (2010) determines that knowledge and comprehension levels are in low order thinking questions, while the rest of levels under higher order thinking questions. Table 2.1 shows the brief definition for each level with illustrative examples.

Table 2.1 Brief definition about for level of Bloom's Taxonomy cognitive level with illustrative examples.

Level Name	Description	Examples
Knowledge Level	Refers to memorizing, remembering, and recalling basic information, facts and terms	• What is a global variable? • Name the artist who painted the Mona Lisa.
Comprehension Level	Refers to student ability to present comprehensive idea and understanding topic based on prior learning by translating, interpreting, organizing, and comparing.	• Compare historical events to contemporary situations • Interpret the pictures.
Application Level	Focuses on student ability to apply gained knowledge to handle new situations	• Apply the rule of law to a new situation • Demonstrate how this could work in an industry setting?

To be continued…

…continuation

Analysis Level	Focuses on breaking down information into components to distinguish, classify, find evidence, assumption and structure, or to determine the relationship between them	• Break down the components of a standard film camera and explain how they interact to make the machine work. • Compare this book to the last book you read.
Synthesis Level	In this level student requires to have the ability to propose a new solution by integrating ideas or/and combining elements together	• Create a new product. Give it a name and plan a marketing campaign. • How can we combine and abstract facts about a software system to create new knowledge?
Evaluation Level	This is the highest order thinking level where the students requires to defend, support, judge or criticize about information or issues according to set of criteria	• How do you think the community should grow or change • Given the data available on a research question, take a position and defend it

The key in writing learning outcomes and questions by Bloom's taxonomy is usually verbs (Kennedy 2006). Therefore, verbs play an important role in Bloom's taxonomy cognitive domain. However, some of these verbs might appear in more than one level, in this case, the context of the question will help to determine the Bloom's level of that question. Figure 2.1 demonstrates some verbs and keywords used in each level in Bloom's Taxonomy cognitive domain.

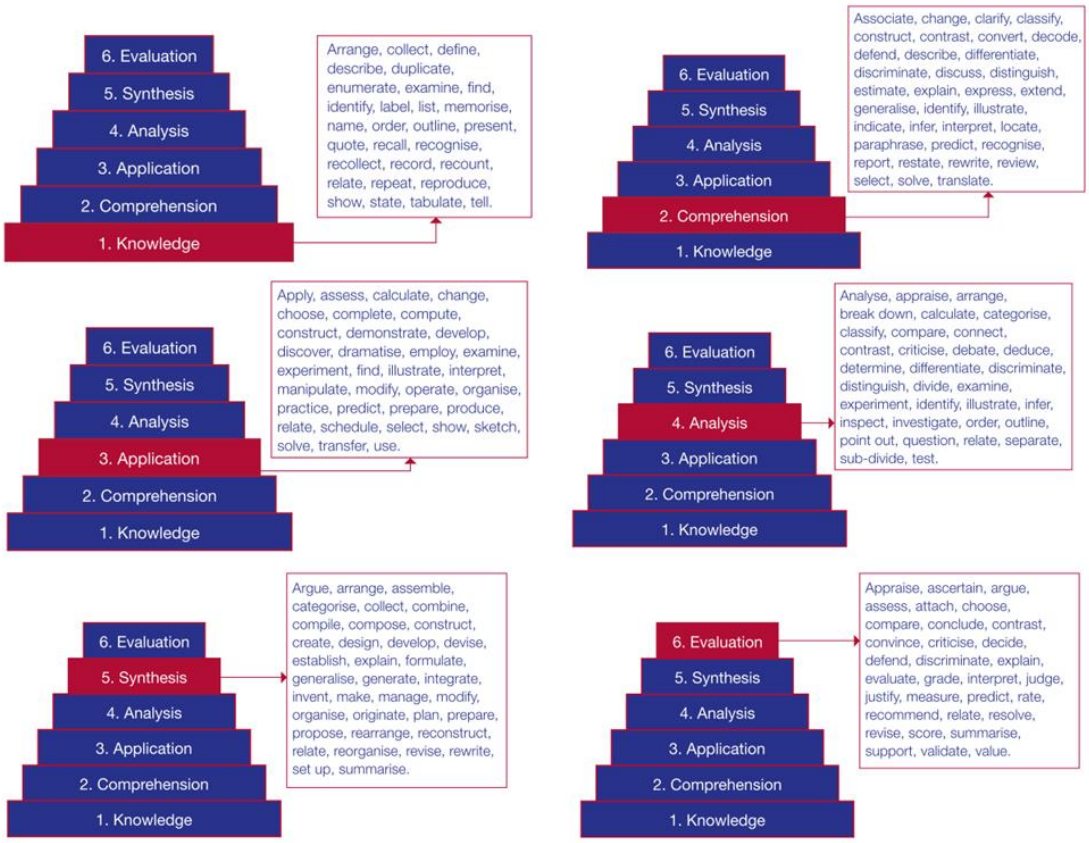

Figure 2.1 Example of verbs and keywords used in each level in Bloom's taxonomy cognitive domain (Kennedy 2006).

In 2001 Anderson & Krathwohl performed and applied some changes to the original Bloom's Taxonomy, in order to make some enhancement and improvement. The improved version called Revised Bloom's Taxonomy. The main difference between the original and the revised version of Bloom's Taxonomy is summarized in three points: emphasis, terminology, and structure. Briefly, the first change is that one more category is added to the knowledge level namely metacognitive category. The second is the names of levels changed from nouns to verbs in order to be more meaningful, descriptive and clear. Knowledge level is named as remember level,

comprehension level is became understand level, application level is changed to apply level, analysis level is called analyze level, create level is a new name of synthesis level, evaluation level is changed to evaluate level. Lastly, the third change is the replacement between the two highest levels, more precisely, among create a level and evaluate level (Krathwohl 2002). Figure 2.2 demonstrates the structure of both versions of taxonomy. This study based on the original Bloom's cognitive domain, since most of the people, are more familiar and prefer the original Bloom's taxonomy (Forehand 2010).

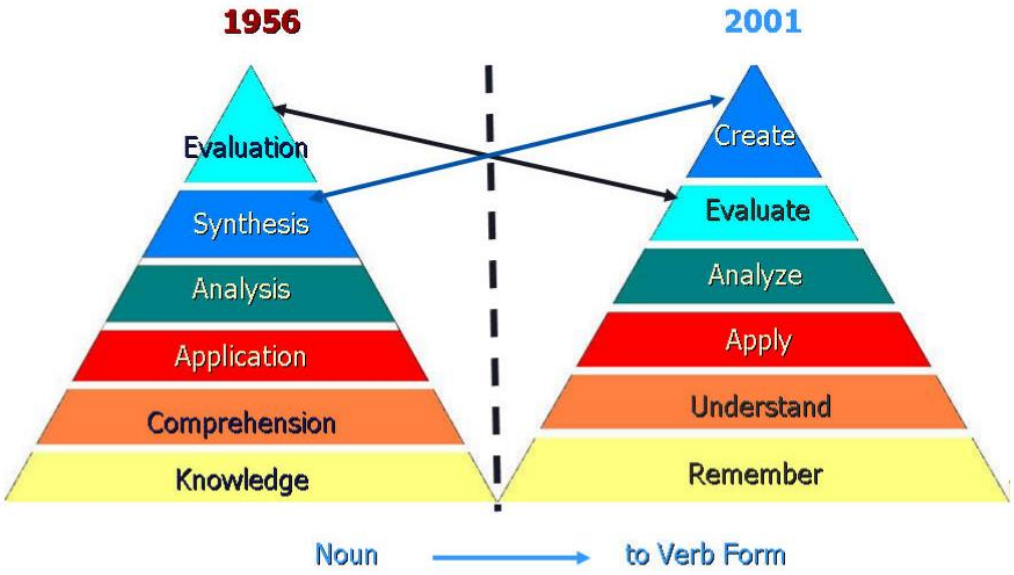

Figure 2.2 Original and revised version of Bloom's Taxonomy (Wilson 2006).

2.3 QUESTION CLASSIFICATION TECHNIQUES

Question classification is a process of assigning questions into suitable predefined classes. It plays an important role in many applications in natural language processing e.g. in auto-generate test and quizzes (Sangodiah et al. 2016; Rasim et al. 2016), and in preparing or evaluating exam papers (Simon et al. 2010; Ginat & Menashe 2015; Jayakodi et al. 2016). Moreover, it is one of the basic building blocks of Question Answering system. Several approaches are used to classify question into appropriate class; rule-based approaches, machine learning approaches, and the hybrid approach as shown in Figure 2.3.

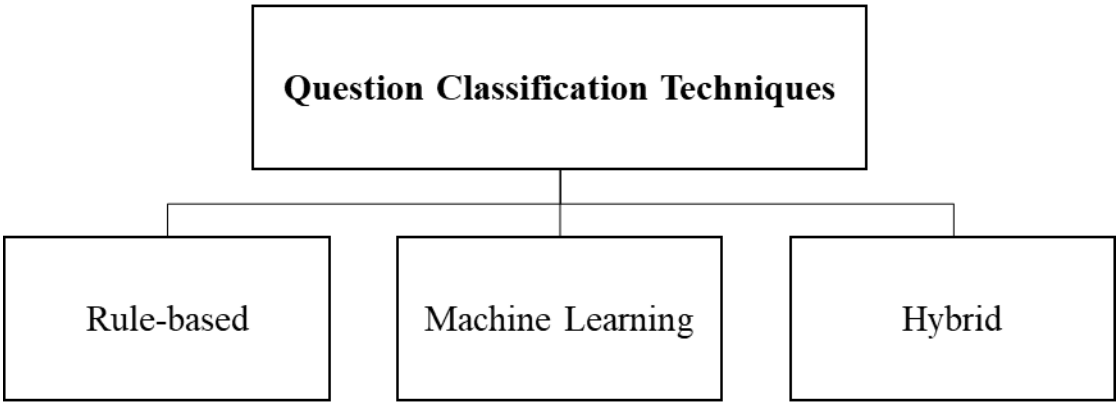

Figure 2.3 Question classification approaches

2.3.1 Rule Based Approaches

The idea behind the rule-based approach is to define a set of rules which is used to match a certain taxonomy. The rule-based approach is a straightforward method. These rules are manually written by experts in a specific domain to handle a specific problem. The question is checked against these rules to map the target class. Several studies in question classification use this approach such as Omar et al. (2012) and Haris & Omar (2015). This approach almost predicts the category of the question accurately, but it requires to write many rules to handle all cases which are exhausted, and time-consuming. In addition, it will perform well with a specific dataset, but not properly with a new dataset. Another drawback of this approach is that it will not be appropriate with a different language and a different domain (Jayalakshmi & Sheshasaayee 2015).

2.3.2 Machine Learning Approaches

Machine learning technique is a suitable solution to tackle the drawbacks of a rule-based approach (Abduljabbar & Omar 2015). It is composed of three kinds of learning; supervised learning, unsupervised learning, and semi-supervised learning. Machine learning manages the process of question classification efficiently, especially the supervised machine learning technique (Van-Tu & Anh-Cuong 2016).

- **Supervised Machine Learning:** one of the most popular techniques used in handling classification problems is Supervised Learning technique. The supervised mechanism involves learning based on a pre-defined set of training examples. In such a way that the training examples are labeled with their corresponding class. Then, the Supervised Learning algorithm uses these training data to distinguish the patterns in each distinct class, these patterns are used to predict the class of unlabeled data i.e. testing data. Many supervised machine learning classifiers are used to classify text, such as Support Vector Machine which performs effectively with unstructured data, according to (Krishnan et al. 2005; Zhiheng et al. 2008; Yahya & Osama 2011; Abduljabbar & Omar 2015; Nirob et al. 2017). In addition, to other classifiers namely; Naïve Bayes, K-Nearest Neighbors, and Artificial Neural Network. Sangodiah et al. (2015) performed a comprehensive review in question classification. The review state that the SVM classifier often outperforms other classifiers since it works well with unstructured text data.

- **Unsupervised Machine Learning:** this mechanism comparing to other is much harder, since the goal of it is to learn patterns without having knowledge about the corresponding class label. Clustering is the most common example of unsupervised technique where the similar instances gather into groups. Paranjpe (2006) proposed an unsupervised approach, clustering method for grouping similar questions together by two steps. The first step is finding the main topic of the question. The second step is to use the lexical and semantic similarity among the questions to the cluster. The outcome of this study shows a promising result.

- **Semi-supervised Learning:** this mechanism is partially supervised. It is appropriate in case there is a lack of labeled data and the rest of data is unlabeled, since the large set of labeled data is not always available. Li et al. (2017) proposed a semi-supervised method with a semantic feature to classify the Chinese questions in the question answering system. The size of the dataset used in this study is 12000 questions, and the number of classes is 15. The dataset divided into 10% testing set, and the rest is the training set. The training set divided into 10% unlabeled questions, and the rest is labeled questions, in order to check the effectiveness of semi-supervised learning. The result demonstrates the ability of the semi-supervised method to handle question classification problem in case of a lack of labeled data.

2.3.3 The Hybrid Approach

A technique that combines both a rule-based approach and a machine learning approach is known as a hybrid approach, which takes the benefit of both approaches. In question classification, a hybrid technique usually uses rule-based approach to extract the headwords or keywords, whereas the machine learning techniques are used to fetch other properties. For example, Sherkat & Farhoodi (2014) proposed a use of combination of rule-based and machine learning as a hybrid technique, to handle question classification task in Persian Question Answering systems. This study handle question from specific domain, and produces a satisfactory result.

2.4 FEATURES IN QUESTION CLASSIFICATION

The classifier cannot understand the text as it is. Therefore, the question should be converted into a vector representation, the format that can be understandable by the classifier. This process is called feature extraction. For the purpose of text classification in general, and question classification specifically, many studies extracted diverse kinds of features with different approaches. The categorization of feature extraction in question classification can be gathered into three distinct classes based on the types of the linguistic information into; lexical, syntactical and semantic features.

2.4.1 Lexical Features

The lexical feature is concerned with the words that exist in a question i.e. word level-feature. N-gram is one of the most commonly used lexical features in question answering and text summarization (Chali et al. 2009). A special case of N-gram is a unigram or what is known also as a bag-of-words, where every single word addressed as a feature. Which count the number of the times the single word appeared in a question. In addition, word shape such as lower case, upper case, and digit is another example of a lexical feature which used widely in question classification (Loni et al. 2011). TF-IDF can be considered as a lexical feature since it is concerned to evaluate the weight of the word in the question.

TF-IDF

Term Frequency- Inverse Document Frequency (TF-IDF) is a statistical feature and corpus-based approach, that is calculated based on the lexical and morphological properties of the text. TF-IDF is a very common weighting method used in Information Retrieval and Text Mining (Chen et al. 2016) which scores the importance of the word in a document. The higher TF-IDF value for the word, the stronger relatedness to the document that appeared in. However, TF-IDF does not handle other information as an effect of word distribution among different classes (Zhu et al. 2016).

Since TF-IDF is one of the most popular weighting terms method, it is used extensively in many studies (Ramos 2003; Xu 2014; Domeniconi et al. 2015; Chen et al. 2016; White 2017), and in question classification (A. A. Yahya & Osama 2011; A. A. Yahya et al. 2012; A. Yahya et al. 2013; Sherkat & Farhoodi 2014; Osman & Yahya 2016; Minai et al. 2018). However, some researchers used TF-IDF as it is, while others proposed some enhancement to the TF-IDF in order to improve the performance.

A Massive Open Online Courses (MOOC) search engine is proposed by (Xu 2014). The methodology of this study based on TF-IDF which assigns the important words in the query by specific weight according to their part of speech. Xu (2014) changed the way of calculating the term frequency by assigning a higher weight for the

most important word which is a noun or verb. Then in the second priority, the adjective and adverb is assigned to the weight that is less than noun and verb but higher than another part of speech which is assigned to the lowest weight. This method produces a significant result. Some other researchers also used the syntactic feature part of speech to assign a suitable weight to the terms such as (Lioma & Blanco 2009; Jovanovska & Zdravkova 2017; Lioma & Blanco 2017). Another example of improving the way of calculating the TF-IDF is proposed by (Zhu et al. 2016) which introduced an impact factor that multiplied by the traditional TF-IDF. The purpose of impact factor equation is to calculate the weights of the class distribution, this enhancement performs well better than the traditional method.

2.4.2 Syntactic Features

Syntactic feature is a feature that can be extracted from the question syntactical structure based on the grammar (Loni et al. 2011; Jayalakshmi & Sheshasaayee 2015). The most popular examples of syntactic features that are widely used in question classification are Part-of-Speech (POS) and question headwords (Van-Tu & Anh-Cuong 2016). A syntactic structure of the question can be represented by the parse tree which based on grammar rules. The parser helps to extract a headword from the question. Part-of-Speech denotes the tag or the class of the word in the question such as Noun (NN), Verb (VB), Adverb (RB), and Adjective (JJ) (Jayalakshmi & Sheshasaayee 2015)

2.4.3 Semantic Features

Semantic feature concerned about the meaning of the term. To extract the semantic features many techniques have been proposed, most of them need a third party source e.g. WordNet (Loni 2000). WordNet dictionary is a lexical database which contains a word's hypernyms, synonyms, and antonyms, that is commonly used in question classification (Loni 2000; Van-Tu & Anh-Cuong 2016). Another semantic feature is word embedding which is also used with many question classification, and text categorization works (Kim 2014; Zhu et al. 2016 W. Zhu et al. 2017; Minai et al. 2018).

2.4.4 Word Embedding

Many natural language processing studies with deep learning models have been included learning word vector representation, where the word vectors are represented in a dense form known as word embedding (Mikolov, Corrado, et al. 2013; Wang et al. 2015; Wang et al. 2016).

Word embeddings also known as distributed representations or word vectors. Word embeddings represent words into dense vectors with low-dimensionality, in which the words that are semantically and syntactically related are closed to each other in the embedding space. Words vectors representation has been used efficiently in many natural language processing tasks (Mikolov, Sutskever, et al. 2013).

Recently word embeddings dominate many conferences proceedings on Empirical Methods in Natural Language Processing and Association for Computational Linguistics. The most popular examples of word embeddings are word2vec, GloVe and fastText. In word2vec, the vectors representation learned via neural-network language model, while in GloVe the vectors representation learned via matrix factorization (Zamani & Croft 2016). FastText similar to word2vec (Mikolov et al. 2017), but the main difference is that word2vec deals with a word as the smallest unit to represent it in vector form. Where fastText treats the word as a bag of character n-grams i.e. subword components, and produce word embedding by summing these subword n-gram. Word2vec developed under Google by (Mikolov, Corrado, et al. 2013). It takes a huge training text as input and establishes vocabulary, then the model learns word vectors representation in such a way that the words share the same context are closed to each other in the vector space. These word vectors can be used as features in various natural language processing tasks. As Kim (2014) mention that, the pre-trained vectors can be considered as universal feature extractors.

Word2vec has two models as shown in Figure 2.4 ; Continuous Bag-of-Word s(CBOW) and Skip-gram. Mikolov, Corrado, et al. (2013) *"The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word"*. However, word vectors capture interesting linguistic

properties semantically and syntactically. As an example for semantic similarity, the words *man* and *woman* are close to each other in vector space, same as *king* and *queen*. In which the distance between *vector(man)* and *vector(king)* is equal to the distance between *vector(woman)* and *vector(queen)*, which can be implemented in simple arithmetic as *vector(king)- vector(man)+ vector(woman) = vector(queen)*. An example for syntactic similarity is *vector (tall)-vector(taller)+vector(short)= vector(shorter)* (Mikolov, Corrado, et al. 2013).

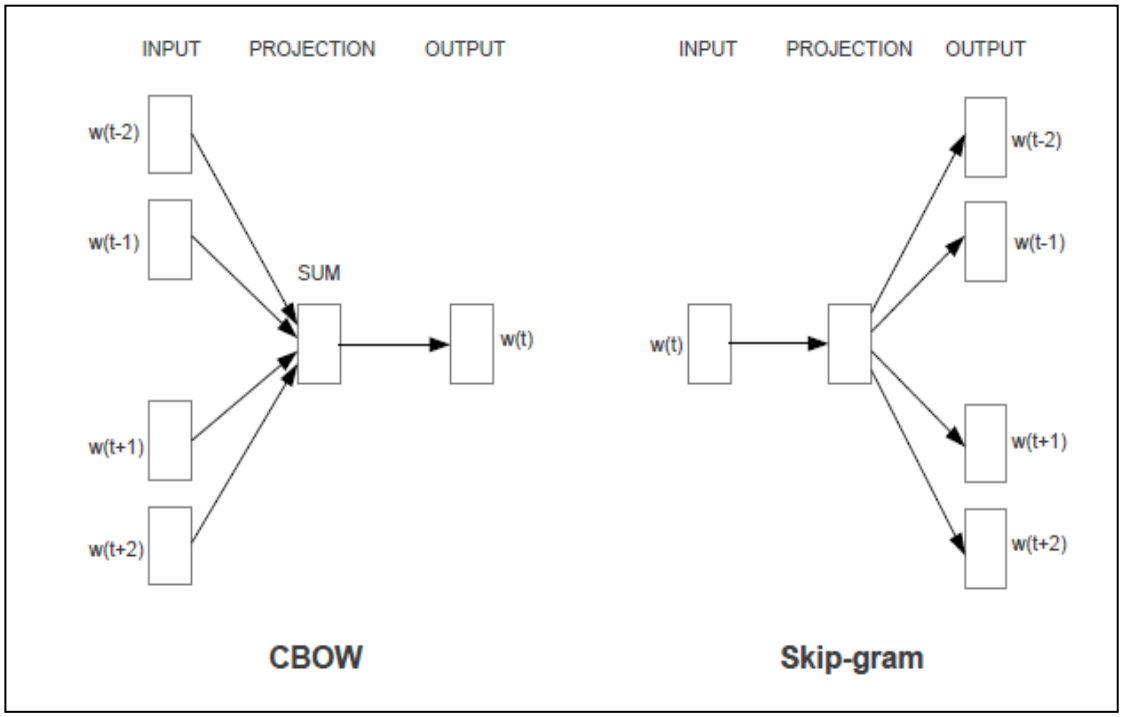

Figure 2.4 Word2vec models (Mikolov, Corrado, et al. 2013)

Kim (2014) and Dahou et al. (2016) stated that, on the absence of large supervised training set, using word vectors that getting from trained unsupervised neural language model to initialize word vectors is a popular way to boost the performance.

Many natural language processing tasks are benefited by word2vec e.g. sentiment analysis, machine translation, and paraphrase detection (Mikolov, Corrado, et al. 2013). Kim (2014) proposed a model that classify question in Question Answering system, and sentiment analysis using Convolutional Neural Network (CNN) that trained on the top of pre-trained word2vec, which produce excellent results. The word2vec used

in this work, has been previously trained by unsupervised neural language model using Google News corpus, which consisted of 100 billion words and produced vectors with 300 dimensions. Similarly, Dahou et al. (2016) used CNN for Arabic sentiment analysis with Arabic pre-trained word2vec which produces a significant result and outperforms previously existed methods.

Another successful result obtained by word2vec is represented in this comparative study (Naili et al. 2017) which used word2vec, GloVe, and LSA as features to perform topic segmentation for Arabic and English dataset. The result state that the word2vec outperforms the other two feature; GloVe and LSA. Moreover, Wohlgenannt et al. (2016) highlighted that the word2vec outperforms GloVe in case of word similarity tasks, where GloVe has superior performance in case of word analogy.

A comprehensive review in question classification by Sangodiah et al. (2015) mention that most of the works in question classification used semantic and syntactic features rather than pure statistical features such as bag-of-word and n-gram. Moreover, Sangodiah et al. (2015) state that semantic features have been significantly used in question classification in question answering system and information retrieval and produced a considerable accuracy, while there is a lack of extracting semantic features in educational environments.

Word2vec with TF-IDF

Since word2vec is the trend nowadays, many researchers (Lilleberg 2015; Zhu et al. 2016; W. Zhu et al. 2017; Minai et al. 2018) use it to handle different natural language processing tasks, along with TF-IDF in order to take the benefit of both features.

Lilleberg (2015) proposed a method that classify text via SVM classifier, with the use of a semantic feature based on word2vec weighted by TF-IDF. In this study, several experiments have been made to compare if TF-IDF is better than combining it with word2vec, whether with or without stop words. Three steps are performed to produce the vector representation of word2vec with TF-IDF. The first step is summing the word2vec vectors to produce single vector. The second step is the summation of