A WORD EMBEDDING METHOD BASED ON K-NEAREST NEIGHBOR FOR HADITH CLASSIFICATION

ALQAHTANI TURKI

PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER IN INFORMATION SYSTEM

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

KAEDAH PENYIRATAN PERKATAAN BERDASARKAN KEJIRANAN K-TERDEKAT UNTUK PENGELASAN HADIS

ALQHTANI TURKI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SISTEM MAKLUMAT

FAKULTI SAINS DAN TEKNOLOGI
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

DECLARATION

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

08 February 2019

ALQAHTANI TURKI
GP06102

ACKNOWLEDGEMENT

First and foremost, I would like to give thanks to the Almighty Allah who has walked with me throughout this journey and before. Without the constant guidance and protection of Allah, this work would barely be a dream.

I would like to express my profound appreciation to my supervisor **DR. Mohd Ridzwan Yakub** for his support, encouragement and continuous motivation during my research. I would like to thank him for sharing his immense knowledge and guiding me to make my thesis better at every step. I appreciate all his contribution of time and finding in making my master experience productive and stimulating.

I am also very grateful to all the teaching and administration staff at Faculty of Information Science & Technology at University Kebangsaan Malaysia who have provided me the opportunity to perform my research in a remarkable scientific environment. I thank them all for their contribution.

Thanks are also given to all my friends and colleagues at the University Kebangsaan Malaysia with the atmosphere created their support and sympathy, I was able to surmount the ordeals and to carry through my thesis.

My special gratitude goes to my father's soul who has provided me spiritual support. Finally, the one person who has made this all possible has been my mother. She has been a constant source of moral and spiritual support and encouragement and has made an untold number of sacrifices for the entire family, and specifically for me to continue my schooling. I received great inspiration from her love and companionship, even from thousands of miles away.

ABSTRACT

Hadith is one of the main legitimate sources for around 1.5 billion Muslims where it considered a guidance for their daily tasks. Several research studies have examined the hadith in the context of text mining. One of the tasks that have been depicted in the literature is the classification of hadith based on its topic or book. Categorizing hadith into its topic or book has been addressed widely in the literature where several methods have been used to accommodate such task. Most of these methods were relying on external knowledge source such as dictionary or lexicon to identify the semantic of hadith text. Although using such sources have improved the classification of hadith however, it requires using massive dictionary that contains comprehensive semantic explanation for words used in hadith. Taking the advantage of recent text representations such as the word embedding, it is not compulsory to use external knowledge source to determine the semantic of words. Word embedding has the ability to give the single word unique embedding based on the context that showed such word. Hence, this study has proposed a new classification method for hadith based on word embedding. In particular, this study has used word embedding to give each word a distinct embedding, additionally, the term frequency and the inverse document frequency TF-IDF have been added to the embedding in order to indicate the importance of the term. Finally, K-nearest neighbor classification method has been used to classify the hadith documents into multiple clusters based on Euclidean distance. A benchmark dataset of Sahih Al-Bukhari has been used in the experiment. Results revealed that the proposed method has obtained an f-measure of 98.37%. Such results demonstrated the effectiveness of the proposed method.

ABSTRAK

Hadis adalah salah satu sumber sah utama bagi seramai lebih kurang 1.5 bilion umat Islam yang menggunakannya untuk tujuan solat dan upacara keagamaan seperti puasa dan mengerjakan haji. Beberapa kajian penyelidikan telah mengkaji hadis dalam konteks pelombongan teks. Salah satu daripada tugas-tugas yang telah digambarkan dalam kesusasteraan adalah klasifikasi hadis berdasarkan topik atau buku. Mengategorikan hadis mengikut topik atau bukunya telah dirujuk secara meluas di dalam kesusasteraan di mana beberapa kaedah telah digunakan untuk membantu kerja sebegini. Sebahagian besar daripada kaedah ini bergantung kepada sumber pengetahuan luar seperti kamus atau leksikon untuk mengenal pasti semantik teks hadis. Walaupun penggunaan sumber tersebut telah memperbaiki klasifikasi hadis, ia memerlukan penggunaan kamus besar yang mengandungi penjelasan semantik untuk perkataan-perkataan yang digunakan di dalam hadis. Dengan mengambil kesempatan gambaran teks yang terbaru seperti penyiratan perkataan, tidak wajib menggunakan sumber pengetahuan luaran untuk menentukan semantik kata-kata. Penyiratan perkataan mempunyai keupayaan untuk memberikan kata tunggal penyiratan unik berdasarkan pada konteks yang menunjukkan perkataan itu. Oleh itu, kajian ini telah mencadangkan kaedah klasifikasi baru untuk hadis berdasarkan perkataan penyiratan. Khususnya, kajian ini telah menggunakan penyiratan perkataan untuk memberi setiap perkataan suatu penyiratan yang jelas, tambahan pula, kekerapan istilah dan kekerapan dokumen songsang TF-IDF telah ditambah kepada penyiratan untuk menunjukkan kepentingan istilah tersebut. Akhirnya, kaedah klasifikasi tetangga K-terdekat telah digunakan untuk mengelaskan dokumen hadis ke dalam pelbagai kelompok berdasarkan jarak Euclidean. Dataset penanda aras Sahih Al-Bukhari telah digunakan dalam eksperimen. Keputusan menunjukkan bahawa kaedah yang dicadangkan telah memperolehi ukuran-f 98.37%. Hasil itu menunjukkan keberkesanan kaedah yang dicadangkan.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 BACKGROUND

Hadith is the second legitimate source for one and half billion Muslims after the Quran in which the hadith is representing the details about how a regular Muslim can perform his duties that have been mentioned in the Quran (Burton 1994). In addition, hadith is also organizing the Muslim individual daily life events and the transactions among Muslims themselves. There are wide range of hadiths which can be categorized into three main classes; Sahih (fully-authentic), Hasan (Semi-authentic) and Da'if (non-authentic) (Bimba et al. 2015). The first class which is the Sahih refers to the hadiths that have been passed from authentic and unbroken chain of narrators whom had a strong memory. Second class which is the Hasan refers to the hadiths that have been passed from authentic and unbroken chain of narrators yet, such narrators might had some limitations regarding recalling specific details. Third class which is the Da'if refers to the hadiths that suffers either from weak memory narrators or a broken chain of narrators.

On the other hand, another taxonomy of hadiths based on the topics in which some hadiths are related to Salat (praying), Haj (pilgrimage) or other topics that might be related to the Muslim worship or the Muslim daily transactions. With the dramatic development on computer technologies, it is an essential demand nowadays to employ such technologies in a way that serve Muslims interactions with hadiths. Therefore, several researchers have exploited the Data Mining and Text Mining techniques in order to extract meaningful patterns from these hadiths. Some authors have examined the questions answering task based on hadiths (Sheker et al. 2016). In this manner, the user can type a question regarding hadith, then the system will respond with the exact

answer. Some researchers have examined the information retrieval based on hadith where the user may type a query regarding some hadith's topic and the system will respond with the relevant hadiths (Harrag et al. 2008). Some researchers have examined the task of Named Entity Recognition (NER) based on hadith in which the proper nouns within hadiths such as persons' names, locations' names and dates are being extracted (Harrag et al. 2011b). Finally, some researchers have examine the Ontology, which is a tree of knowledge where taxonomies such as worship would contain several branches including praying, pilgrimage and fasting (Baraka & Dalloul 2014; Dalloul 2013).

However, the most common and vital task regarding hadith mining is the classification of hadiths in terms of its topic. This task is essential for the former tasks that have been discussed earlier. This is due to classifying the hadiths based on topic will help other tasks to figure out the degree of authenticity of such hadith, identifying its narrators and indexing these hadiths into multiple topic ranks. With the emergence of Machine Learning Techniques (MLTs) which are statistical methods that intended to build a model for classification, the hadith classification has caught many researchers' attentions (Saloot et al. 2016). However, MLTs are mainly relying on the features used within the dimension space.

Within the dimension space of hadith classification into its relevant topic, there are several features that could be used including linguistics, statistics and distributional. Linguistic features refer to the semantic characteristics that can be used to identify the synonym of terms in which a term such as 'Salat' would be associated with terms like 'zuhur', 'aser' or 'maghrib'. Apparently, this type of features require using an external knowledge source such as dictionary, thesaurus or ontology in order to identify the synonymy of terms. On the other hand, statistical features refer to the number of occurrence for each term within a hadith and how such term could represent significance indication to a particular topic. Finally, distributional features refer to the characteristics that could be used for the terms distribution among the hadith. Specific measures such as Mutual Information can be used to identify the distribution of a particular word in accordance to the hadiths.

In this regard, this study aims to utilize a new representation for the terms within hadith

in order to improve the hadith classification.

1.2 PROBLEM STATEMENT

Several research studies have examined the hadith classification in which the hadith is being classified into its accurate topic. For example, Jbara (2010) have used a classification based on linguistic features where stemming and word expansion have been used to facilitate the classification. Stemming refers to the task of retrieving the word's root such as *fasting* and *fast*, while word expansion refers to the task of expanding a term with its synonyms such as *Salat* and *praying*. Apparently, this method requires using a knowledge dictionary which is not always available for use especially for Arabic language. Several studies have followed such way of classification in which an external knowledge source is being used to guide the classification such as (Alkhatib 2010; Baraka & Dalloul 2014; Harrag 2011).

With the emergence of more sophisticated representations such as the Word-to-Vector, the use of external knowledge source has become unnecessary. Word to vector representation aims to give each term an embedding based on the context. It is working on a single layer neural network to provide a unique and distinct embedding for every term. Even the single term would have multiple embedding based on the context. For example, the word 'Salat' (i.e. praying) might be occurred as one of the main five prayers such as 'Salat zuhur' and 'Salat maghrib' or it might be occurred as a special occasion such as 'Salat al-mayyit' (i.e. prayer for death) which is being accommodated on a person who has been passed away. Word-to-vector representation has the ability to give different embedding for the word 'Salat' based on its context.

The advantage of using such representation is that it can discard the use of external knowledge source to identify the contexts of terms, meanwhile, producing more robust and accurate classification accuracy. However, word-to-vector cannot address statistical information of the words such as the frequent occurrence in accordance to a particular document.

Examining the statistical information of the word would facilitate to identify the most

accurate set of features. Since the word will represent the features thus, it is necessary to determine the most appropriate set of words that has a significant impact on the classification accuracy. Therefore, this study aims to examine the Term Frequency Inverse Document Frequency along with the word-to-vector in order to identify the best features.

1.3 RESEARCH OBJECTIVES

The objectives of this research can be summarized as follows:

i. To propose a feature selection approach based on Term Frequency – Inverse Document Frequency (TF-IDF) and word embedding for improving hadith classification.

ii. To apply K-Nearest Neighbour classification method based on the feature selection proposed in objective (i).

1.4 RESEARCH MOTIVATION

The main motivation behind this study is the great potential shown in the literature for the Word-To-Vector representation. The last decade has shown numerous research study in the field of text mining where the use of external knowledge source was vital. Determining the synonymy of terms within the text classification was performed using such external knowledge source. For this purpose, researchers were spending their efforts whether on finding an existing dictionary or affording to build a new dictionary. Even after building such dictionary, there are numerous terms that could be missed or misinterpreted. Therefore, replacing the use of dictionary or any external knowledge source with a sophisticated method that can avoid building the dictionary, and at the same time, providing similar classification performance, was an imperative and challenging task.

Word-To-Vector has been articulated to do such task in which the neural network architecture can be used to provide a unique and distinct embedding for each term within the text corpus. Such embedding not only treating the individual terms but also treating the multiple meanings of each terms. This can be represented by giving different embedding for the single term based on the context.

Another motivation can be represented by the need of determining the most appropriate set of features (i.e. words). Basically. Each word within the word embedding has its own impact on the classification accuracy. Some of them have significant impact while others have insignificant. Therefore, there is a vital demand to identify the most accurate set of words. Taking the advantage of Term Frequency – Inverse Document Frequency (TF-IDF), determining the significance of each word can be gained by attaching a weight of TF-IDF within the word embedding to indicate the significance of words.

1.5 RESEARCH SCOPE

This study aims to propose a Word-To-Vector representation method for the hadith classification in Arabic language. Such classification aims to categorize each hadith into its relevant topic. For this purpose, Sahih-Albukhari has been used as text corpus that contains variety of hadiths. In addition, several pre-processing tasks are being used to filter each hadith from the unwanted information such as stopwords and numbers. Furthermore, a single layer neural network architecture is being used to process each term and outputs its embedding. In addition, TF-IDF will be used as a feature selection approach where each embedding of the words will be supplemented with a TF-IDF weight to indicate the importance of the word.

Finally, a K-Nearest Neighbour (KNN) classification method has been used to classify the hadith into its topic. The reason behind using such classifier lies on its ability to deal with the real numbers. Since the word embedding consists of series of real numbers therefore, the most suitable classifier would be used is the KNN.

The evaluation of the classification method will be performed using the traditional machine learning assessment metrics including precision, recall and f-measure. These metrics are intended to evaluate the correctness of predictions provided by the classifier.

1.6 RESEARCH METHODOLOGY

The methodology of this study consists of four main stages including (i) Problem Identification, (ii) Design, (iii) Implementation, and (iv) Evaluation as shown in Figure 1.1. The first stage aims to identify the problem statement of this study. This can be performed by reviewing the literature of hadith classification and determine the existing limitations and gaps. Second stage aims to design an appropriate solution for the defined gap in the previous stage. Identifying a suitable solution would require also reviewing the literature regarding the latest methods used for improving the classification task.

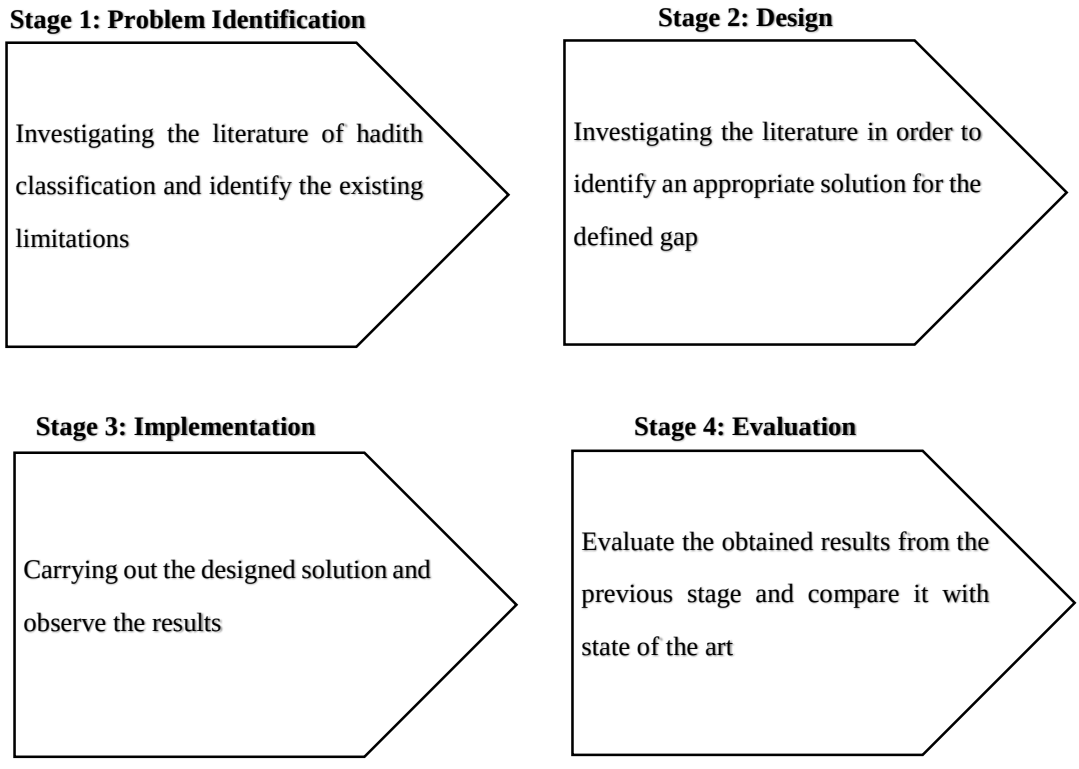

Figure 1.1 Research methodology stages

Third stage aims to apply the designed solution that has been determined in the previous stage. Such application requires determining an appropriate dataset for the

experiment, as well as, observing the results of these experiments. Finally, the fourth stage aims to evaluate the obtained results from the experiments. Such evaluation will be based on specific metrics such as precision, recall and f-measure. In addition, a comparison with the state of the art would be conducted in this stage in order to clarify the novelty of the proposed solution.

1.7 RESEARCH ORGANIZATION

This research is organized into five chapters in which each chapter is tackling a specific aspect of the research methodology. These chapters can be illustrated as follows:

Chapter I provides the basic components of each research where the background of study is being discussed along with the problem and objectives. In addition, the methodology of the study is being discussed in this chapter.

Chapter II discusses in detail the literature review behind the hadith classification in which the classification methods are being declared. In addition, the type of features and representations that usually used with the classification task are also illustrated by determining its strength and weak points. In addition, the related work in hadith classification will be highlighted. Finally, a critical analysis of the related work will be provided in this chapter. Based on such critical analysis, the gap of the study will be declared.

Chapter III discusses in detail the procedures of applying the proposed method in which the dataset used in the experiments is being declared. In addition, the pre-processing tasks that have been conducted are also discussed in this chapter. Furthermore, the feature extraction and selection will be discussed in this chapter. After that, the classification process is being described. Finally, the evaluation method that is being followed by this study will be described.

Chapter IV highlights the experimental results obtained by the proposed method where the correctness of KNN prediction is being evaluated. In addition, the experiment settings that have been used to gain the results are being discussed in this chapter. In

order to show the novelty of KNN, further experiments will be depicted in this chapter using other classification methods. Finally, a comparison against the state of the art is being conducted to clarify the novelty of the proposed method.

Chapter V provides the final summary of the research in which the final conclusion is being illustrated. Furthermore, the contribution of this study will be discussed in this chapter. Finally, the future work that could be motivated by this study will be tackled in this chapter.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter aims to discuss the literature review behind hadith classification. In order to provide a comprehensive discussion, a brief background for the hadith mining tasks has been provided in Section 2.2. In this section, all the factors that are significantly affecting the hadith mining process are being identified. Then, each factor has been discussed separately in an independent section. Therefore, Section 2.3 discusses several purposed of hadith mining. Section 2.4 illustrates the features that could be used within the hadith classification process. Section 2.5 focuses on the possible feature selection tasks that could be used to improve the hadith classification. Section 2.6 highlights the learning topologies that could be used to accommodate the hadith classification. Section 2.7 provides a discussion for common classification algorithms. Finally, Section 2.8 concentrates on the hadith classification and the previous studies that tackle this task.

2.2 BACKGROUND OF HADITH MINING

Hadith is one of the main sources for approximately one and a half billion Muslims whom using such source for organizing their daily life activities along with their worships. Therefore, researchers have addressed the hadith in different tasks such as hadith's topic classification, hadith's authentication or so-called 'إسناد / isnad' classification, named entity recognition of hadith narrators, and extracting knowledge from hadith. Most of these tasks are belonging to a broader area of study which is known as machine learning techniques.

Machine learning techniques are one of the data mining tasks that have been extensively

examined in many areas of studies such as information retrieval and information extraction (Tan 1999). The aim of machine learning techniques can be expressed as the task of predicting a class label for a set of records based on a historical data that has been annotated with a class label (Gilad-Bachrach et al. 2004). Applying machine learning techniques for hadith in the literature took several forms in which various aspects are being involved. Such aspects includes purpose of hadith mining, features used and learning topology.

First of all, different purposes have been be used within mining hadith such as the document classification, named entity recognition, and document analysis (Saloot et al. 2016). Second, the features used within the hadith mining is another aspect that has a significant impact on the whole process (Gopal & Yang 2010). Features can be defined as the discriminative characteristics that facilitates the process of identifying the class label. There are numerous type of features that could be used in which it is highly affected by the purpose of hadith mining. Third, the topology of learning during hadith mining resembles another factor that is significantly affecting the classification task. There are three main topologies of learning topology including supervised, semi-supervised and unsupervised. Each topology indicates the extent of supervision which refers to the process of training on the historical data.

All the aforementioned factors that are playing essential role during the hadith mining, will be tackled in further details within this chapter. Following sections will discuss each of these factors independently.

2.3 PURPOSE OF HADITH MINING

In fact, there are data mining tasks that have been examined for the hadith domain. The first task was the hadith document classification. This task has been addressed earlier for classifying the hadith document into Sahih (i.e. fully-authentic), Hasan (i.e. Semi-authentic) and Da'if (i.e. non-authentic) according to the degree of hadith authentication (Bimba et al. 2015). Table 2.1 shows a sample of hadith documents along with their class label of their degree of authenticity.

Table 2.1 Example of classifying hadith based on authenticity

Data	Class
Hadith Document 1 *The Prophet said: "Do not tell a lie against me for whoever tells a lie against me (intentionally) then he will surely enter the Hell-fire."* *قال رسول الله: "لا تكذبوا علي فمن كذب علي متعمدا فليتبوأ مقعده من النار"*	Non-Authentic ضعيف
Hadith Document 2 *The prophet said: "Know that Paradise is under the shade of the swords"* *قال رسول الله: "واعلموا أن الجنة تحت ظلال السيوف"*	Semi-Authentic حسن
Hadith Document 3 *The Prophet said: "A Muslim is the one who avoids harming Muslims with his tongue and hands* *قال رسول الله: "المسلم هو من سلم المسلمون من لسانه و يده"*	Fully-Authentic صحيح

As shown in Table 2.1, the data contains multiple hadith text document where each of them consists of massive text words. Based on the context inside these documents, the classification task is intended to categorize these hadith into multiple classes based on authenticity (i.e. fully-authentic, semi-authentic and non-authentic hadith). Dalloul (2013) and Baraka & Dalloul (2014) have extensively examined the hadith in terms of authenticity by analyzing both the context of the hadith and the narrators (i.e. isnad).

Another task that have been depicted in the literature regarding the hadith domain is the classification based on topic. In fact, hadiths are numerous and belong to a wide range of topics which are arranged in multiple books such as ‘the book of belief’, ‘the book of revelation’, and others. Therefore, researchers have analyze the context of the hadiths in order to classify them into their actual book or topic. Table 2.2 shows a sample of hadith documents along with their corresponding class labels of actual topic or book.

Table 2.2 Example of classifying hadith based on topic or book

Data	Class
Hadith Document 1 *The Prophet said: "None of you should offer prayer in a single garment that does not cover the shoulders."* *قال رسول الله: " لا يصلي أحدكم في الثوب الواحد ليس على عاتقيه شيء "*	Book of praying كتاب الصلاة
Hadith Document 2 *The Prophet said: "The prayer of a person who does ,Hadath (passes, urine, stool or wind) is not accepted till he performs (repeats) the ablution."* *قال رسول الله: "لا تقبل صلاة أحدكم إذا أحدث حتى يتوضأ"*	Book of Ablutions (Wudu') كتاب الوضوء
Hadith Document 3 *The Prophet said: "A Muslim is the one who avoids harming Muslims with his tongue and hands* *قال رسول الله: "المسلم هو من سلم المسلمون من لسانه و يده"*	Book of belief كتاب الايمان

As shown in Table 2.2, every hadith document has been classified into specific topic or book based on the context of the hadith itself. Harrag (2011) has examined the classification of hadith based on its topic by using Sahih Al-Bukhari book of fully authentic hadiths.

The last task of hadith mining that have been addressed in the literature is the named entity recognition. This task is intended to identify the proper nouns such as people's name, location's name, organization's name, dates and times (Nadeau & Sekine 2007; Steinberger & Pouliquen 2007). In this regard, the narrators of the hadith will be examined in this task. Table 2.3 shows a sample of classifying narrators' names within the hadith of:

"According to Saed Bin Alhareth, he saied, we asked Jabir Bin Abdullah about the praying in a single cloth \ "عن سعيد بن الحارث قال سألنا جابر بن عبدالله عن الصلاة في ثوب واحد

Table 2.3 Example of named entity recognition with hadith

Data	Class
عن \ according to	O
سعيد \ Saed	Narrator
بن \ bin	Narrator
الحارث \ Alhareth	Narrator
قال \ he said	O
سألنا \ we asked	O
جابر \ Jabir	Narrator
بن \ Bin	Narrator
عبدالله \ Abdullah	Narrator
عن \ about	O
الصلاة \ the praying	O
في \ in	O
ثوب \ a single	O
واحد \ cloth	O

As shown in Table 2.3, the task of named entity recognition transforms the text into a series of words rather than sentences in which each word is being examined in terms of proper nouns or not. The name 'سعيد \ Saed' has been classified as a narrator. Harrag et al. (2011b) and Siddiqui et al. (2014) have examined the task of named entity recognition to extract the narrators' names within the hadiths.

2.4 FEATURE TYPES FOR HADITH MINING

This section aims to discuss the type of features that can be used for different types of hadith mining tasks. Basically, features play a significant role in the classification where the classifier cannot predict the class label of an instance data using only the raw data and its corresponding class label in the historical data or training data. It should be some significant characteristics that may indicate the desired class label (Blum & Langley 1997). Assuming a data of hadith documents that need to be classified based on topic, it is necessary to consider the features within the hadith itself in order to classify it to its topic. In this regard, there are main types of features that could be used within the hadith analysis including Distributional features and Knowledge-based features. Following sub-sections will address these two type of features.

2.4.1 Distributional Features

This type of features aims to examine the distribution of terms within the hadith text. Considering the hadith document classification where the document is being classified into its topic, it is necessary to identify what type of terms used in each document. In this regard, the distributional features are representing the terms that occurred in particular document in order to facilitate determining the class label (Wang et al. 2013). This type of feature is knowns as N-gram or Bag-of-words (BoW).

In order to understand the N-gram features that could be used, let a dataset of hadith document as depicted in Table 2.4.

Table 2.4 Sample of dataset of hadith documents

ID	Arabic Hadith	Translation
Document 1	*قال رسول الله: " لا يصلي أحدكم في الثوب الواحد ليس على عاتقيه شيء "*	*The Prophet said: "None of you should offer prayer in a single garment that does not cover the shoulders."*
Document 2	*قال رسول الله: "لا تقبل صلاة أحدكم إذا أحدث حتى يتوضأ"*	*The Prophet said: "The prayer of a person who does ,Hadath (passes, urine, stool or wind) is not accepted till he performs (repeats) the ablution."*
Document 3	*قال رسول الله: "المسلم هو من سلم المسلمون من لسانه و يده"*	*The Prophet said: "A Muslim is the one who avoids harming Muslims with his tongue and hands*

Now in order to use the N-gram features within the hadith documents in Table 2.4, it is necessary to examine all the terms included in the three hadith documents as attributes. Table 2.5 shows such organization of attributes for N-gram terms.