Developing A novel Text mining model for exploring knowledge from An Arabic text

Abstract

- The Arabic language seen as its rich and complex language. Root extraction is one of the most important topics in the context of natural language processing applications. The formation of the Arabic word depends on abstraction, which is the root of the word.
- Stemming is an attempt to reduce word to its root form. It is a pre-processing step in Text Mining Applications as well as a very common need thing of Natural Language Processing (NLP) filed, Information Retrieval systems and text classifiers.
- In this study, we developed a new text mining model by using visual basic language. The model consist of two algorithm. We applied our model on (Sahih Al-Bukhari) textbook.
- We achieved 95.6% root extraction accuracy and 95.5% inflection accuracy.
- Our model built a Hadith dataset to become a source for any exploring knowledge to assist us for solving life problems.

Problem Statement

- We found a few studies work on exploring knowledge from the unstructured Arabic text corpus, because the complexity of the Arabic language in terms of the syntactic structure due to the diacritics signs that affects to the grammatical form of the Arabic word, which gives us more than one form and different meaning when we compared to the English and French languages, which considered one of the challenges facing researchers now a days.
- We develop a text mining model with a new technique relied on the Arabic diacritics text to extract the proper root and its inflection for each word with high accuracy to exploring the knowledge by determining the entities in Al-Hadeeth Al-shareef corpus as verb tenses (Past, Present, and Imperative), noun and proper noun with high performance.

Motivations

- We find that a few studies have dealt with the Arabic text and Al Hadeeth Al-shareef, most of them dealt with extracting word root or inflection, classification and confidence of the narrators and analysis in the text of the hadith.
- The aim of our research is to enrichment the Arabic language by developing text mining algorithms to exploring knowledge from Arabic corpus. Hadith as the second source in Islamic legislation.
- In this research, we will build a complete text-mining model for exploring knowledge from Al Hadith Al-shareef corpus.

Contribution

- The main contribution of this study can be display as follow:

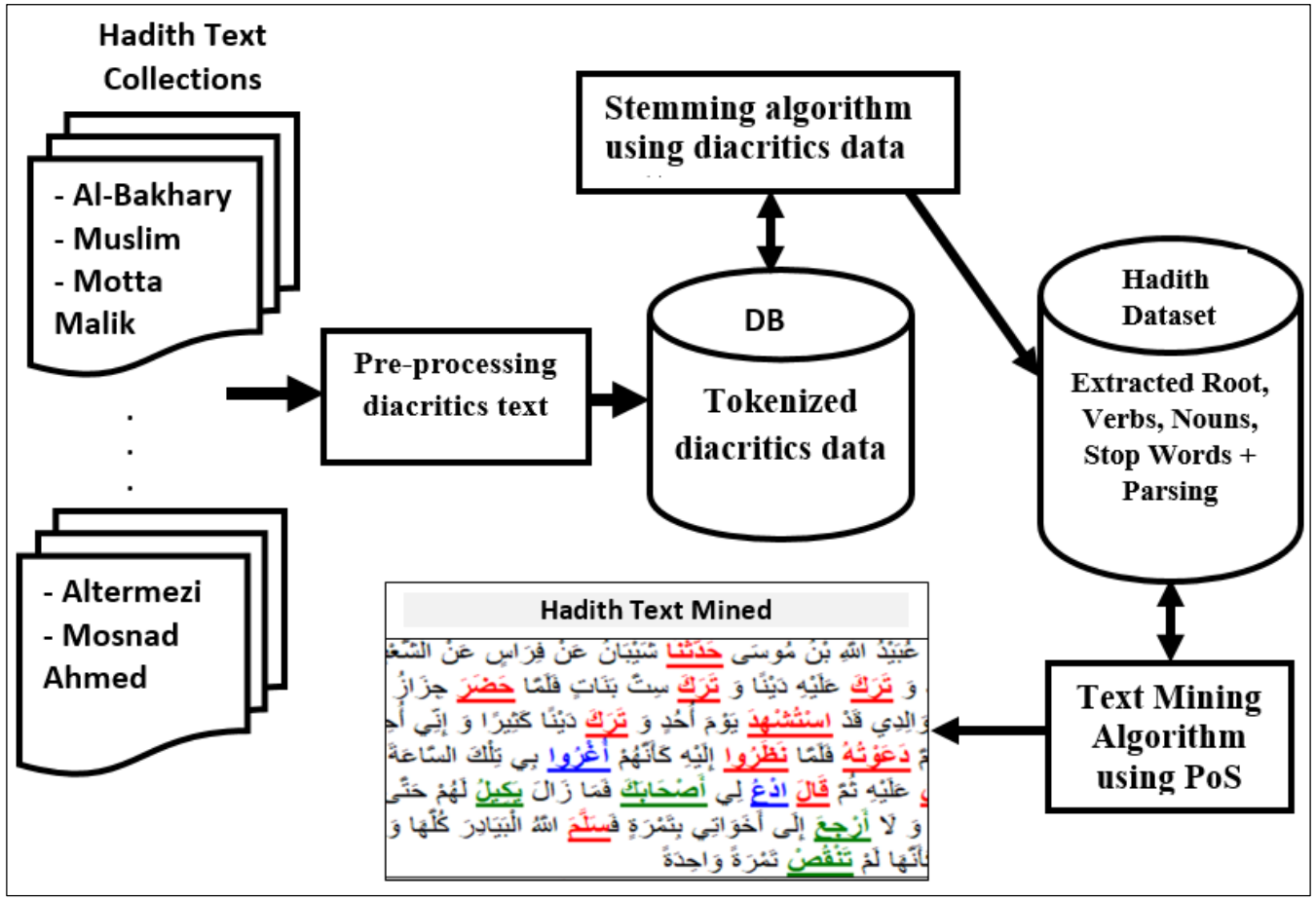
  Built a new text mining model with new concept for:

  - ❑ extracting word root with accuracy 95.6% (IR).
  - ❑Find word inflection with accuracy 95.5% (classification)
  - ❑Knowledge Exploration by determine entities(verb tenses, nouns) in Al Hadeeth Al-Shareef.

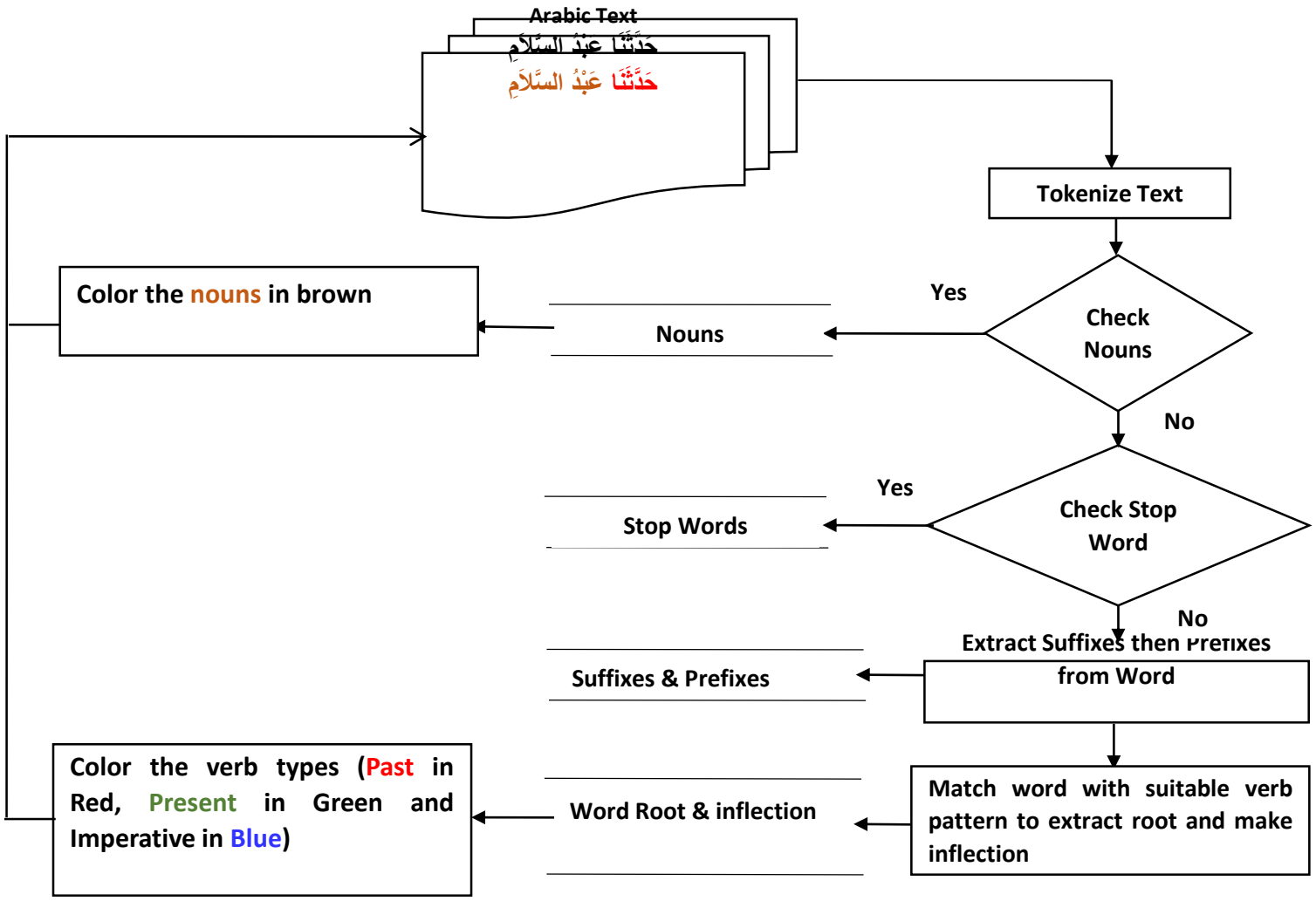
Source Data

- In this research we construct data set for Sahih Al-Bukhari Book by collecting from Hard and soft copy of Sahih Al-Bukhari book and Web sites

Proposed Model



Proposed Stemmer and text mining algorithms



Result

- In our research, we built knowledge base for Hadith it can used in multiple applications in Hadith classification and text mining approaches and other machine learning applications
- The challenge of this study is building a model to improve the search effectiveness of knowledge extraction and Arabic Information Retrieval (AIR)
- If the users search for something they can applied their query by using keyword in a different formats in any search engine. The new developed stemming algorithm capable to cluster all same forms of the word that have the different meaning to different group based on their semantic.

Model implementation

- The root extraction and parsing algorithm receives the Arabic text of the hadith of the Prophet as shown in figure bellow:

Invocations:كتاب الدعوات Chapter: Invocation during the Sala:tباب الدُّعَاءِ فِي الصَّلاَةِ	
Narrated `Abdullah bin `Amr: Abu Bakr As-Siddiq said to the Prophet, "Teach me an invocation with which I may invoke (Allah) in my prayer." The Prophet (ﷺ) said, "Say: Allahumma inni zalamtu nafsi zulman kathiran wala yaghfirudhdhunuba illa anta, Faghfirli maghfiratan min indika war-hamni, innaka antalGhafur-Rahim."	حَدَّثَنَا عَبْدُ اللَّهِ بْنُ يُوسُفَ، أَخْبَرَنَا اللَّيْثُ، قَالَ حَدَّثَنِي يَزِيدُ، عَنْ أَبِي الْخَيْرِ، عَنْ عَبْدِ اللَّهِ بْنِ عَمْرٍو، عَنْ أَبِي بَكْرٍ الصِّدِّيقِ ـ رضى الله عنه ـ أَنَّهُ قَالَ لِلنَّبِيِّ صلى الله عليه وسلم عَلِّمْنِي دُعَاءً أَدْعُو بِهِ فِي صَلاَتِي. قَالَ " قُلِ اللَّهُمَّ إِنِّي ظَلَمْتُ نَفْسِي ظُلْمًا كَثِيرًا، وَلاَ يَغْفِرُ الذُّنُوبَ إِلاَّ أَنْتَ، فَاغْفِرْ لِي مَغْفِرَةً مِنْ عِنْدِكَ، وَارْحَمْنِي، إِنَّكَ أَنْتَ الْغَفُورُ الرَّحِيمُ ". وَقَالَ عَمْرٌو عَنْ يَزِيدَ، عَنْ أَبِي الْخَيْرِ، إِنَّهُ سَمِعَ عَبْدَ اللَّهِ بْنَ عَمْرٍو، قَالَ أَبُو بَكْرٍ ـ رضى الله عنه ـ لِلنَّبِيِّ صلى الله عليه وسلم.
Reference :Sahih al-Bukhari 6326	

Model Implementation Cont.

- processed all the words of the hadith, starting by identifying nouns and stop words, then removing the suffixes and prefixes, and then putting them in the appropriate pattern to extract the word root and then performing the inflection process to determine the verb tenses as shown in the following table:

DocNo	HadeethNo	WordNo	OrginWord	ProcessedWord	Pattern	WordTypeEng	WordTypeAr	Root	VerbTime	Preffix	Suffix
78	6326	1	حَدَّثَنَا	حَدَّثَ	فَعَّلَ	Verb	فعل	حدث	فعل ماضي(past verb)		نَا
78	6326	2	عَبْدُ	عَبْدُ		Stop Word	كلمة توقيف				
78	6326	3	اللَّهِ	اللَّهِ		Noun	إسم				
78	6326	4	بْنُ	بْنُ		Stop Word	كلمة توقيف				
78	6326	5	يُوسُفَ	يُوسُفَ		Noun	إسم				
78	6326	6	أَخْبَرَنَا	أَخْبَرَ	أَفْعَلَ	Verb	فعل	خبر	فعل ماضي(past verb)		نَا
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
78	6326	51	ارْحَمْنِي	ارْحَمْ	افْعَلْ	Verb	فعل	رحم	فعل أمر(imperative verb)		نِي
78	6326	52	إِنَّكَ	إِنَّكَ		Stop Word	كلمة توقيف				
78	6326	53	الْغَفُورُ	الْغَفُورُ		Noun	إسم				
78	6326	54	الرَّحِيمُ	الرَّحِيمُ		Noun	إسم				
78	6326	55	عَمْرُو	عَمْرُو		Noun	إسم				
78	6326	56	الْحَارِثِ	الْحَارِثِ		Noun	إسم				
78	6326	57	يَزِيدَ	يَزِيدَ		Noun	إسم				
78	6326	58	إِنَّهُ	إِنَّهُ		Stop Word	كلمة توقيف				
78	6326	59	سَمِعَ	سَمِعَ	فَعِلَ	Verb	فعل	سمع	فعل ماضي(past verb)		
78	6326	60	عَبْدَ	عَبْدَ		Stop Word	كلمة توقيف				
78	6326	61	بْنَ	بْنَ		Stop Word	كلمة توقيف				
78	6326	62	أَبُو	أَبُو		Stop Word	كلمة توقيف				
78	6326	63	سَلَّمَ	سَلَّمَ	فَعَّلَ	Verb	فعل	سلم	فعل ماضي(past verb)		

Thanks