# The Classification of Persian Texts with Statistical Approach and Extracting Keywords and Admissible Dataset

Ehsan Mohtashami
Department of Computer,
Science and Research Branch,
Islamic Azad University,
Bushehr, Iran

Mehrnoosh Bazrafkan
Department of Computer,
Science and Research Branch,
Islamic Azad University,
Bushehr, Iran

## ABSTRACT
In recent years, a lot of algorithms have been proposed for the classification of the documents. Most of works done have been on English language and recently there have been works on some languages such as Chinese, Arabic, etc. In some cases, there were classifications on the Persian texts which have become essays or online projects.

One of the algorithms that have been used most frequently in text Classification is KNN algorithm which is more frequently in the texts Classification in the English language. In order to use these algorithms we need suitable dataset of Persian texts, which unfortunately these data are not available to Persian Texts Classification .So the our first and second phase in this project are extracting the keywords and creating Admissible dataset for the classification of the Persian texts, and The third phase of this project is to implementing a software for the classification of the Persian texts using the extracted keywords.

In this essay, we have reviewed and paid attention to some challenges of searching and classifying the Persian texts, and we have also implemented an application in order to extract the admissible dataset for the classification of the Persian texts with statistical approach or with KNN and N-gram and etc, which produces some suitable and usable dataset for the classification of the Persian texts. In the last phase we have also implemented an application in order to classify the Persian texts with a statistical approach.

## General Terms
Natural Language Processing, Text Classification

## Keywords
Persian Texts Classification, Statistical Classification, extracting Persian texts keywords, extracting dataset

## 1. INTRODUCTION
Primary purpose of this article is classification Persian texts to the nine classes (economic, political, events, art, sport, literary, medicine, historical and religious). Previously these actions have been done on English language. A good measures has been done in this regard on Persian language is the activities of "Noor text mining group" that a tools of the online text classification is prepared to users.

The articles has been worked about Persian texts Classification, often have used the KNN algorithm. To implement the classifier of Persian texts we needs to Keywords in Persian language and Persian dataset extracted in this area, but unfortunately any dataset were not available. So we started to extract keywords and acceptable dataset and implement an application to classify Persian text. [2]

## 2. TEXT CLASSIFICATION

### 2.1 The importance of text categorization

There are large amount of textual information on the Internet and in different computing systems various firms. Classification of this information manually is time consuming and expensive and there is a need specialist groups that are available 24 hours a day. So the automatic software to classification thousands documents in a day without the experts is great Helpful.

### 2.2 Data Collection

Data set that we have chosen to classify Persian texts is a Part of the Text corpus of the Hamshahri newspaper in different years that as form of XML files were available But because little correspondence with our target groups, we tested classification on this data and on different data that were extracted from news.

Hamshahri textual corpus data is suitable for text processing and data mining and among the works that have been done on it can note to "Hamshahri of Tehran University". Of course, there are problems such as XML tags and the old characters and Arabic characters such as "ى" and "ک" can be seen in this collection.[3]

### 2.3 Searching and Classification Problems in Persian Texts.

In addition to problems that to exist in searching and classification Persian texts, there are good ways that sometimes makes work easier. Including that to search for number of family and similar words in various forms, we can use a word or the root for search all those words.

Below refers to some of the problems of search and classification Persian Texts:

#### *2.3.1 There Are Two Different Symbols for The Letters "ى" And "ک"*

One of the major problems in Persian language on the Internet is different spelling letters. This means that for each of the two letters "ک" and "ى", there are two different spellings, in better Expression; for each of them there are two different

characters. For example, for the letter "ک", there are two characters with Unicode 1705 and 1603 that codes 'ک' in Persian and "ك " in Arabic. In the last letter of the Persian alphabet, there are two characters with Unicode code 1740 and 1610 to the Code, "ی" Persian "ي" Arabic.

If the keywords and the text search, from these terms are different, the results are not accurate.

### *2.3.2 Words That Are Similar in Character And different In Meaning and Subject*

Consider the word "تصادف" in Persian, if you search the word "تصادف" and want to be classified, can be placed in the "accidents" class and "art" Class, Because the word "تصادف" ends with "دف" that is a other word means One musical instruments, So it can be placed in the "Art" class. The same can be said about the pair of words "مادر" and "ماد" is also.

To solve this problem, for example, we can introduce the word "دف" in keywords with spaces before and after the word. [1]

### *2.3.3 Words That Can Be Placed In Two Classes*

Consider the word "داور", which can correctly be two categories, "Sport" and "art".

For example, there are the words, "حمله" (sporting and events), "آمریکا" (political, scientific, medical) and "حزب الله" (religious, political) too.

As much as possible keywords should be tried in every category, those are specific words.

### *2.3.4 Time-Limited Words*

Some words are limited to a time or a period. For example, the President's name can be seen in several years as a keyword in Political Class and after that time may lose their status.

### *2.3.5 Texts That Can Be Placed In Two Categories*

There are texts that contain the actual properties of two separate classes, without having to do the wrong software classifier. For example, consider the text on the PC market and the price of parts and new technologies and increasing price that can be placed on both scientific and economic.

To resolve this problem, giving weight to certain words that have more value, would be a suitable way.

### *2.3.6 Two Words Phrases*

Phrases such as "آتش سوزی" (آتشسوزی), "سپاس گزار" (سپاسگزار) and plural names such as "زمین ها" (زمینها) and "توپ ها" (توپها), etc, which are written in two ways: separately and sticking together. Of course, spelling, separately writing is correct, but in the old texts and some new texts can be found.

# 3. IMPLEMENTATION

## 3.1 Extract Persian Keywords

To extract dataset and classification of Persian texts in 9 categories (economic, political, events, arts, sports, literature, medicine, history and religion), we need keywords in this categories. We extract about 45 to 60 keywords for each category and then store them as standard orthography in different modes such as vector, batch and combination.

Important points to choose keywords:

1) Try to select keywords among the unique terminology and specific names.
2) Be avoided common words that included many categories.
3) Be considered completely, personal name or name of the object or place to search.

## 3.2 Extracting Dataset of Persian Keywords

The second phase is to Extracting dataset of Persian keywords that is used for subsequent works such as classification with algorithms KNN (see Figure 1).

We extract dataset using the keywords specified in a way that comes: First, we store all the keywords in a vector whose length is the number of words (451) and then we search for all the 451 words in all of texts. We get the number of occurrences of all keywords in any of the texts, with the search.

We did the searching with the software that we implemented with C# and the result of search is 451 element vectors that are stored in a TXT file sequentially and respectively, the number of texts. We want to get the number of occurrences of each keyword in texts.

For example, consider searching the extracted keywords in a XML news collection of the Hamshahri newspaper has 167 Separate texts. The output of this step is 167 vectors which have 451 elements and each element represents the number of occurrences of the keyword in the text searched.

Also on this level the application store a TXT file which shows a real class of the entire Texts.

The output of this phase (dataset extraction) is the three following files:

**Dataset.txt**: The number of occurrences of 451 keywords in entire texts searched.

**Class.txt:** Name or ID of all the texts along with their real class.

**Lexicon.xlsx:** Includes a 451-elements vector and keywords that will determine the order of words and their class.

## 3.3- Implementing an Application to Classify the Persian Texts, Using Statistical Approach

The implementation this phase has been done with C# too, and in the final was prepared an application to classification Persian texts.

With this software, users can give any text to the application and the software will determine the class or subject of the text (see Figure 2).

This software has been tested with 167 texts from News collection of Hamshahri newspaper and 100 texts New News and in testing of 100 texts from News which had certain topic, the application classified the texts With an accuracy of 90%, correctly.

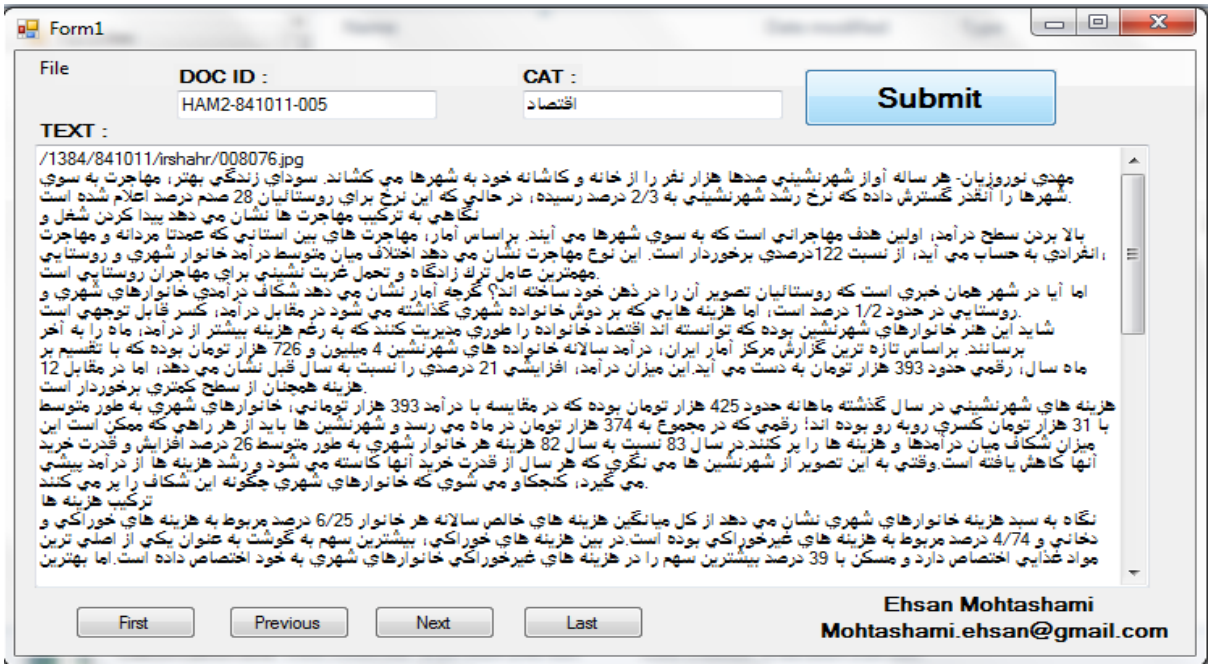


**Fig 1: The Screen Shot of the software which Designed for Extracting Dataset**

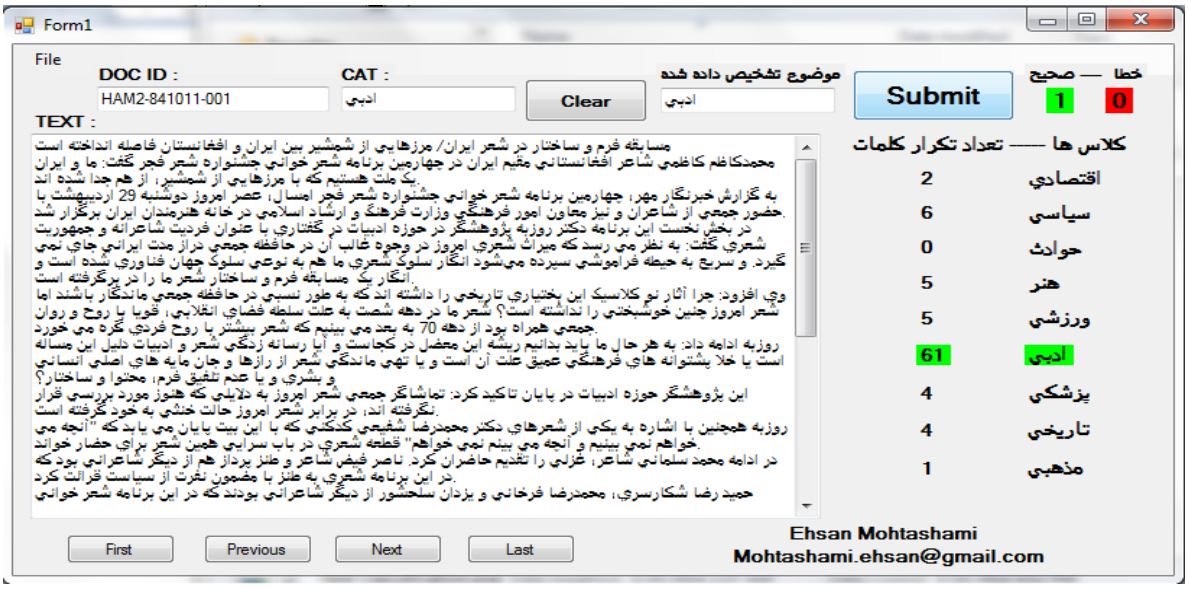


**Fig 2: The Screen Shot of the software which Designed for classify the Persian texts.**

## 4. CONCLUSION

Extracted Dataset and application of this paper have been evaluated. Extracted Dataset can be used to working with KNN algorithms and N-gram for Persian texts classification and the application designed also works well, and classify Persian texts with accuracy of 90%.

## 5. FUTURE WORKS

At the end of the third phase, we obtained very useful information on classify Farsi texts and also can be used from the extracted dataset in the later works and Persian texts classification with KNN and N-gram algorithms.

This software is suitable for extraction dataset from any set of documents and users can use it in your own projects.

It is hoped that the next version, the software has higher accuracy and be available to users in the form of offline and online.

## 6. REFERENCES

[1] Andrew Roberts , January 2009, Grammatical Inference and Corpus Linguistics , Submitted in accordance with the requirements for the degree of Master of Philosophy, The University of Leeds School of Computing.

[2] Shahla Nemati, Mohammad Ehsan Basiri, Persian documents classification using KNN algorithm.

[3] The HAMSHAHRI collection of Tehran University: www.ece.ut.ac.ir / dbrg / hamshahri

[4] The Website of Noor text mining group is : www.textmining.noorsoft.org