An Approach For Inspecting Arabic Authorship Attribution Using The Fuzzy Logic System

Abeer H. El Bakly, Nagy Ramadan Darwish and Hesham A.Hefny

Abstract: using fuzzy logic system enhances the accuracy of solving the society problems so, the researchers try to use it in many research issues such as authorship attribution specifies an author to an anonymous text according to his writings. The type of features is character, lexical, syntactic, and semantic. This research studies the effect of using fuzzy logic system on the Arabic authorship attribution within proposing a new approach to solve the issue of detecting if an anonymous text is owning to specific author or not. Additionally, it uses the ontology as a semantic feature and N-gram as a lexical feature. Moreover it concerns on a new corpus which is called “ElWafaa LiFokahaa”. It is assembled and annotated fatwas of prayer from books of main Islamic jurisprudence doctrines (“Hanfi, Shafie, Malik, and Hanbli”). In the experimental work, the corpus is split into two datasets, 95% is training and 5% testing dataset. In addition, the proposed approach uses the semantic and lexical feature with and without fuzzy logic system then compare them. The implementation shows the effect of fuzzy logic system on the semantic and lexical feature. The accuracy is reduced from 96.7% to 94.6% with semantic features while it is increased from 91.6% to 95.7% with lexical features.

Index Terms: Arabic Authorship Attribution, Artificial Intelligence, Fuzzy Logic System, Machine Learning, Lexical Feature, N-gram, ontology, semantic feature.

1 INTRODUCTION

The (FLS) is a model to modeleize the human capacity of uncertainty reasoning. In the fuzzy logic all realities are fractional, it is termed the specific parameter between extremes of a yes and no which is considered the degree of the truth. Using FLS in solving the problems increases the near to real world, so this research uses FLS in authorship attribution (AA) research domain [3]. This research domain involved many issues such as assigning anonymous text to the specific author of the set authors according to his writing features. The features of the author reflect the sexiness, age, religion, training, position, inspiration or philosophy [8]. The lexical feature presents the text into groups of tokens which are term, numeral, or a punctuation mark. Semantic features are produced from dependency schemas that include two types of data: binary semantic features and semantic adjustment connections [15] [5]. This research uses the ontology as semantic feature. Ontology is part of artificial intelligence, knowledge system and information system. Ontology includes the relations between different sets of terms in a specific scope. It provides kind of object, notion, features and relations [20]. In the Islamic community, people need knowing if a particular fatwa is possessed to the particular Islamic doctrine or not. This research proposed a new approach for detecting who constitutes the fatwa of the main Islamic doctrines by using FLS with ontology as semantic feature and n-gram as a lexical feature. It may be summarized the contributions of this research as below: Using FLS with Arabic ontologies with different classes as a semantic feature and N-gram as lexical feature in Arabic AA. Analysis the effect of using FLS on the semantic and lexical feature then comparing the both of features. Establishing a novel dataset “ElWafaa LiFokahaa” that is assembled and annotated the fatwas of the prayer for main Islamic jurisprudence doctrines from Islamic books manually.

The rest of the research is as follows: part 2, introduces the background about the fuzzy logic system, the definition of AA; the definition of ontology, part 3, introduces the related works which include issues of AA; part 4, introduces a proposed approach; part 5, analyses an experimental work and results, part 6, clarifies the conclusion and future work.

2 BACKGROUND

2.1 Fuzzy Logic System

Fuzzy logic system (FLS) can be characterized as the nonlinear mapping of the facts to a scalar yield information. Figure 1 shows the architecture of an FLS.

![Fig. 1. The architecture of fuzzy logic system](image_url)

An FLS consists of four main parts as below:

2.1.1 Fuzzification Process

The processing of fuzzification data is conversion each part of the input data to degrees of the membership by specific membership functions. Membership functions can be symmetrical or asymmetrical. They are characterized on one-dimensional or multidimensional universes. There are diverse forms of membership functions such as: Triangular and Trapezoidal.

---

*Abbeer H. El Bakly, Faculty of Graduate Studies for Statistical Research, Cairo University. Email: abbeerhassan012@gmail.com*

*Nagy Ramadan Darwish, Information system and technology department, Faculty of Graduate Studies for Statistical Research, Cairo University. Email: nagyr@cu.edu.eg*

*Hesham A.Hefny, computer system department, Faculty of Graduate Studies for Statistical Research, Cairo University.*
2.1.2 Fuzzy Rule-Based system
The Rule-Based System contains a set of “if-then rules” which reflect the set of facts or information. These “if-then rules” usually are used with fuzzy logic which can deal with complex real-world issues, handling the uncertainty. In addition, using fuzzy logic make inference methods more flexible with approximate reasoning methods. These “if-then rules” formulate the conditional statements whose depend on antecedents and consequent of Fuzzy Logic [16][17].

2.1.3 Defuzzification [4]
It is defined as a procedure of transferring fuzzy values to crisp values. Defuzzification is the method of transferring a fuzzified yield into a one crisp value with deference to a fuzzy set. In addition, it is processed according to the yield parameters of the membership functions. The common methods mostly are used for defuzzification are listed as below:
- Center of Sums Method (COS)
- Center of gravity (COG) / Centroid of Area (COA) Method
- Center of Area / Bisector of Area Method (BOA)
- Weighted Average Method
- Maxima Methods

2.2 Definition of Authorship Attribution
Figure 4 presents the architecture to solve the AA research issue which contains a group of points as that: The principal point is a corpus or dataset which might be a group of essays or stories or documents … etc. It reverberates the features of the author. The next point is the extraction of the features from the author’s text. These features have various kinds for instance lexical features (sequential terms that are assembled into phrases, character features (sequential characters), syntactic features and semantic features. There are two methods for characterizing these features of the dataset [9][2]. The first method is profile-based that merges the author’s documents in the single file. The output of extraction the single file singular vector of features. The second method is instance-based which get the author’s documents as one by one so the output of extraction the multiple files is a group of vectors. The next point could use or not the feature selection methods to decrease the features to efficient features. The last step of the architecture is using the similarity techniques such as Euclidean Distance, The Jaccard’s Co-efficient, The Dice’s Co-efficient and Cosine Similarity [9][2].

![Triangular membership function](image1)
![Trapezoid membership function](image2)

![Fig. 3. Trapezoid membership function [10](image3)

![Fig. 4. The architecture of AA with extraction in main two attribution dataset methods](image4)
3 RELATED WORK
Arabic AA, recently Arabic researchers used AA for many purposes. Omar et al [6] proposed a novel stylometric approach which was established on morphological patterns and letter mapping characteristics. This approach assumed unique and distinct stylistic features for each author to discover the authors of disputed texts. They used a dataset included 259 Arabic flash fiction romances with four authors. They used the cluster analysis to group texts which had participated linguistic characteristics. They terminated that morphological data was used in the useful way to enhance the performance of AA. In addition, they enhanced revealing in Arabic texts according to the individual stylistic characteristics of the affixation operation in Arabic. The Arabic argumentative texts could be owned to their authors according to revealing morphological patterns with trustworthy performance. Another purpose for introducing the solution of AA research problem is to study the impact of increasing the size of the training dataset on performance of the classifiers. So AL-Sarem et al [7] classified the dataset in groups training (TS1: the training dataset with 330 documents, TS2: the training dataset with 378 documents, TS3: the training set with 414 documents, TS4: the training set with 438 documents) which were consisted of 4631 texts documents. They composed the features N-gram, Part-Of-Speech and arranged in various groups. The dataset was the text of Islamic fatwas from 1896 to 1996 which was extracted from “Dar Al-ifta AL Misriyyah” website. They grouped fatwas by the word length and size into different groups. They used these classifiers (The Mahalanobis distance (MD), Linear Regression (LR), and Multilayer Perceptron (MP)). The result showed that the composition of POS with N-grams increased the accuracy while the composition of WLF, N-grams, and POS reduced the accuracy. El Bakly et al [8] proposed a novel model joined the three orientations in related work. This model utilized Arabic ontology as semantic feature to detect the anonymous text to the specific author. The dataset was Islamic fatwa for traveller’s prayer of main Islamic jurisprudence doctrines which involved 1073 fatwa (the training set was 751 fatwa, the testing set was 322 fatwa). They utilized the protégé to construct four Arabic ontologies with the same classes. They utilized the KNIME platform for next points in the proposed model specifically measuring cosine similarity to calculate the closeness among anonymous text and known text. They evaluated the proposed model by calculating the right answer of the model over the total queries which gave 90% in their experiments.

4 THE PROPOSED APPROACH
The main issue for this paper is specifying the anonymous fatwa to specific doctrine from the main Islamic jurisprudence doctrines (“Hanfi, Malik, Shafie, and Hanbili”). This paper resolves this issue by proposing an approach utilizing a fuzzy logic system with N-gram as a lexical feature. Using a fuzzy logic system is the pioneer in AA which is concerned the top contribution of this paper. The other one is building a new corpus that contains the Prayer fatwas of main Islamic doctrines.

4.1 The Details Steps of the Proposed Approach
Figure 5 presents the major phases of the proposed approach as below: The principle phase presents the new corpus which is prepared in the .csv file for using in KNIME. Then it is split to training dataset (95%) for each doctrine (“Hanfi, Malik, Shafie and Hanbili”) that is used in the second phase and test dataset (5%) which is used in the third phase. The second phase utilized the training dataset as input then utilizing the loop to extract the features from every doctrine. The parts of this phase as below: The principle part presents the pre-processing tools for processing text such as the Arabic token tool for tokenization, snowball stemmer for stemming the words, and the number filter. The next part, extraction features from the processed text by a bag of words which converts text to words then term frequency computes the various occurrences of words then computing IDF-TF which is used a score of representing each feature in each fatwa. The final part utilized the document vector for converting text to vectors. The output is four vectors, every vector is assigned to every doctrine (“Hanfi, Malik, Shafie, and Hanbili”) with the same sort as figure 5. The third phase utilized the testing dataset (anonymous fatwas) as input then utilizing the pre-processing tools such as the Arabic token tool for tokenization, snowball stemmer for stemming the words, and the number filter. Then extraction the features from the processed text by a bag of words which converts text to words then term frequency computes the various occurrences of words then computing IDF-TF which is used a score of representing each feature in each query. The output is a vector of features that is utilized as a query by using the row filter. The fourth phase includes two parts, the first one is without using fuzzy logic system and the other one is with using fuzzy logic system. Without using FLS, using a loop for inspecting the similarity between each vector of features (query) and each vector of feature for each doctrine. Until this point find the topmost crisp score of the queries which are used in the sixth phase to compute the accuracy for the proposed approach without fuzzy logic system. With Using FLS, this process includes a set of steps as below: Fuzzifying used the input data which is produced from the first part of this phase, it is the similarity each vector of features (query) and each vector of feature for each doctrine. Converting the crisp scores to the fuzzy logic scores by membership functions. Defuzzifying the fuzzy logic scores to crisp scores then find the topmost score for the queries. The fifth phase presents the maximum crisp scores with or without FLS to get the answer which is name of doctrine. The sixth phase contains an assessment for the proposed approach by computing the accuracy.

Fig. 5. The proposed approach for assigning anonymous fatwa to the particular doctrine with and without fuzzy logic system
4.2 Membership Functions of Input Values

Description of each membership function for each cosine similarity between the queries in test dataset and training dataset.
Cosine similarity ("Hanbli"): The fuzzy scores are represented by the membership functions which are triangle and trapezoid. The triangle presents lowest and highest areas of the similarity, medium area of the similarity which is presented by trapezoid membership function. It is illustrated in Figure 6.

Cosine similarity ("Hanfi"): The fuzzy scores are represented by membership functions which are triangle and trapezoid, triangle presents lowest and highest areas of the similarity, medium area of the similarity which is presented by trapezoid membership function. It is illustrated in Figure 7.

Cosine similarity ("Malki"): The fuzzy scores are represented by membership functions which are triangle and trapezoid, triangle presents lowest and highest areas of the similarity, medium area of the similarity which is presented by trapezoid membership function. It is illustrated in Figure 8.

Cosine similarity ("Shafi"): The fuzzy scores are represented by membership functions which are triangle and trapezoid, triangle presents lowest and highest areas of the similarity, medium area of the similarity which is presented by trapezoid membership function. It is illustrated in Figure 9.

For this input it is possible to write: \( \text{Hanfi} = \{l, m, h\} = \{1, 0.5, 0\} \). This procedure is applied to all inputs to get the fuzzy scores for all doctrines.

5 EXPERIMENTAL WORK AND RESULTS
In the experiments of the proposed approach, this research depends on the platform which is called KNIME [18].

5.1 The Specifics of the Corpus
This corpus involves 5477 fatwa (1242 fatwas for “Hanfi”, 1358 fatwas for “Malki”, 1511 fatwas for “Shafie” and 1366 fatwas for “Hanbli”) which is called “ElWafaa LlFokahaa”. The extraction of fatwas is executed from the books of prayer which is part of worship in Islamic jurisprudence doctrines [11][12][13][14] etc which were downloaded from the website of el maktba el shamla(http://shamela.ws/). Moreover, each researcher can access and download this corpus freely from this link https://studcu-my.sharepoint.com/:f:/g/personal/abeerhassan_pg_cu_edu_eg/Evi0lAWDPApPmFNSTGr78CoBTRBdlswsXd8Il-QiN0mCQ?e=kGAIx7 [19]. Table3 clarifies the details of this corpus.
Table 2
The specifics of “ElWafaa LFokahaa” corpus

<table>
<thead>
<tr>
<th>Doctrine</th>
<th>No. fatwas</th>
<th>No. terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanfi</td>
<td>1242</td>
<td>45088</td>
</tr>
<tr>
<td>Malik</td>
<td>1358</td>
<td>33620</td>
</tr>
<tr>
<td>Shafie</td>
<td>1511</td>
<td>29708</td>
</tr>
<tr>
<td>Hanbli</td>
<td>1366</td>
<td>41546</td>
</tr>
</tbody>
</table>

Figure 11 introduces a sample of fatwas for a doctrine that is the Hanfi’s view of his doctrine. It contains the condition to get the license of the travel prayer that defines the objective of travel which should have a good or bad as the highlight red circle.

Figure 11. The fatwa for Hanfi’s doctrine

5.2 Steps of Pre-Treatment Corpus
The pre-treatment of the corpus is executed in the proposed approach by utilizing KNIME in the below points:

- Tokenization: each fatwa is splintered in phrases then each phrase is splintered into tokens (small units) for instances words or characters.
- Stemming: each base of word is extracted to stem or root word for instance ("يقصر", "قصر", "قصير") to "قصر".
- Number filter: the used Arabic language is Standard Arabic so deleting numbers is mandatory.

Table 3
The example of processed sentence in different features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sentence representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original sentence</td>
<td>من شروط السفر لا يكون ممكناً &quot;مئن شرط السفر لا يكون ممكنا&quot;</td>
</tr>
<tr>
<td>Word Unigrams</td>
<td>&quot;مئن&quot;,&quot;شرط&quot;,&quot;السفر&quot;,&quot;لا&quot;,&quot; يكون&quot;,&quot;ممكنا&quot;</td>
</tr>
<tr>
<td>Word Bigrams</td>
<td>&quot;مئن&quot;, &quot;شرط&quot;, &quot;السفر&quot;, &quot;لا&quot;, &quot;يكون&quot;, &quot;ممكنا&quot;</td>
</tr>
</tbody>
</table>

5.3 N (1, 2, 3, 4, 5)-gram as Lexical Feature
This research proposes using the word unigram, word bigram, word trigram, word fourgram and word fivegram.

Table 3 presents an example for applying the previous features. The method of applying these features as below:

- Applying the proposed approach with each feature.
- Getting the values of cosine similarity between each doctrine in training dataset and test dataset.
- Calculating the average value for five features.
- Applied the rest of proposed approach.

5.4 The Extraction Process Details of Features
Figure 12 introduces the handling text in the .cvs files by utilizing KNIME at the top part of the figure. Every file is clarified as numeral vectors of features by utilizing some techniques such as BOW as a basic stage. BOW is a popular technique for extraction the features to convert the text to single terms. Then computing the TF and the IDF then computing the TF-IDF for the processed file by the math formula red square at the low side of the figure 12 which presents the extraction stages. The last stage is introduced as the last highlighted box in the figure is document vector to convert all processed data to vector of feature. The figure 13 presents the details of stages for extraction features.

5.5 The detecting process with and without fuzzy logic
Table 4 presents the part of 274 queries (anonymous fatwas) as 5% of the corpus which is test dataset, the proposed approach is applied with KNIME for every query then the answer value is the doctrine’s name. The specifics of table 4 are as below:

Table 4
The specifics of “ElWafaa LFokahaa” corpus

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sentence representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original sentence</td>
<td>من شروط السفر لا يكون ممكناً &quot;مئن شرط السفر لا يكون ممكنا&quot;</td>
</tr>
<tr>
<td>Word Unigrams</td>
<td>&quot;مئن&quot;, &quot;شرط&quot;, &quot;السفر&quot;, &quot;لا&quot;, &quot;يكون&quot;, &quot;ممكنا&quot;</td>
</tr>
<tr>
<td>Word Bigrams</td>
<td>&quot;مئن&quot;, &quot;شرط&quot;, &quot;السفر&quot;, &quot;لا&quot;, &quot;يكون&quot;, &quot;ممكنا&quot;</td>
</tr>
</tbody>
</table>

Fig. 12. The processed of feature extraction in KNIME
The principle column (No. Query) introduces the query number. The next column (“Hanbli”) introduces summary of the cosine similarity between anonymous fatwa (query) and Hanbli’s doctrine which is computed as the average for each part of lexical feature N-(1, 2, 3, 4, 5) gram severally. The next column (“Hanfi”) introduces summary of the cosine similarity between anonymous fatwa (query) and Hanfi’s doctrine which is computed as the average for each part of lexical feature N-(1, 2, 3, 4, 5) gram severally. The next column (“Malki”) introduces summary of the cosine similarity between anonymous fatwa (query) and Malki’s doctrine which is computed as the average for each part of lexical feature N-(1, 2, 3, 4, 5) gram severally. The next column (“Shafie”) introduces summary of the cosine similarity between anonymous fatwa (query) and Shafie’s doctrine which is computed as the average for each part of lexical feature N-(1, 2, 3, 4, 5) gram severally. The next column (“query answer value”) introduces the topmost value of main Islamic doctrines, the next column (“query answer”) introduces the name of doctrine. The next column (“query answer status”) introduces the value of the query answer status which gets two values true or false according to the view of a domain expert in Islamic jurisprudence doctrines.

Table 5 presents the output of applying the fuzzy logic system in the output of the table 4 which are considered the crisp scores for each doctrine that is get as input for membership functions for each doctrine as figures (6, 7, 8, 9). The next step uses weighted average method as defuzzifying method for each column. The previous steps give the same columns as table 4 with the different values with the previous steps.

**TABLE 5**

<table>
<thead>
<tr>
<th>No. Query</th>
<th>Hanbli</th>
<th>Hanfi</th>
<th>Malki</th>
<th>Shafie</th>
<th>query answer value</th>
<th>query answer</th>
<th>query answer status</th>
</tr>
</thead>
<tbody>
<tr>
<td>q1</td>
<td>0.621717</td>
<td>0.580813</td>
<td>0.74896</td>
<td>0.572643</td>
<td>0.74896</td>
<td>malk</td>
<td>TRUE</td>
</tr>
<tr>
<td>q2</td>
<td>0.605601</td>
<td>0.585622</td>
<td>0.74004</td>
<td>0.579962</td>
<td>0.74004</td>
<td>malk</td>
<td>TRUE</td>
</tr>
<tr>
<td>q3</td>
<td>0.586056</td>
<td>0.641993</td>
<td>0.74579</td>
<td>0.620671</td>
<td>0.74579</td>
<td>malk</td>
<td>TRUE</td>
</tr>
<tr>
<td>q4</td>
<td>0.606176</td>
<td>0.569586</td>
<td>0.748416</td>
<td>0.606176</td>
<td>0.748416</td>
<td>malk</td>
<td>TRUE</td>
</tr>
<tr>
<td>q5</td>
<td>0.587182</td>
<td>0.584974</td>
<td>0.748841</td>
<td>0.587182</td>
<td>0.748841</td>
<td>malk</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

5.6 The analysis applied and not applied fuzzy logic in AA

**TABLE 6**

<table>
<thead>
<tr>
<th>No. query</th>
<th>Lexical feature (n-gram)</th>
<th>Lexical feature with fuzzy logic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. right answers</td>
<td>No. wrong answers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>274</td>
<td>251</td>
<td>23</td>
</tr>
</tbody>
</table>
From Table 6 using lexical feature without and with fuzzy logic system in the proposed approach for detecting the fatwa.

The number of queries is 274 which are 5% of the corpus which is considered the test set (anonymous fatwas). The table is split to two parts: The first one describes the results of implementation the proposed approach without fuzzy logic system which are the first column (no. right answer) gives 251 right answer from 274 queries. The second column (no. wrong answers) gives 23 wrong answers then the third column introduces the accuracy which equals 91.6%. The second one describes the results of implementation the proposed approach with fuzzy logic system which are the first column (no. right answer) gives 262 right answer from 274 queries. The second column (no. wrong answers) gives 12 wrong answers then the third column introduces the accuracy which equals 95.7%.The table 6 clarifies that all results of applying the FLS are higher than not applying the FLS.

![Accuracy Graph](image)

**Fig. 14. The accuracy of lexical feature with and without fuzzy logic system**

without and with fuzzy logic system. This accuracy is computed as equation 1.

\[
\text{The accuracy} = \frac{\text{the right answers/ all queries}}{100} = 91.60\%.
\]

The accuracy without FLS = \((251/274)\times100 = 91.60\%\)

The accuracy with FLS = \((262/274)\times100 = 95.70\%\)

From the figure the accuracy with applying FLS is higher than the accuracy without FLS (95.7%, 91.6%). using FLS in the proposed approach which is inspecting who wrote the fatwa from the main Islamic jurisprudence doctrines enhances the accuracy. This means utilizing the FLS with lexical feature enhances the accuracy for inspecting Arabic AA.

6 CONCLUSION AND FUTURE WORK

The fuzzy logic system is used for solving the complex real-world problems, handling the uncertainty. In addition, using fuzzy logic system make inference methods more strong flexible with approximate reasoning methods. There is a big problem in the Islamic society which is knowing who wrote the anonymous text (fatwa) from the known text (main Islamic jurisprudence doctrines). So this research proposed a new approach for solving this problem using fuzzy logic system with N-(1, 2, 3, 4, 5) gram as a lexical feature in the AA that is the basic contribution in this paper. Moreover, this paper introduces a new corpus is called “EIWafaa LiFokahaa” that is another contribution. It is manually assembled and annotated the fatwas of the prayer from books of worship which is part of Islamic jurisprudence doctrines. The KNIME platform is utilized in the next steps of experimental works. The experimental works consist of implementing the proposed approach with and without the fuzzy logic system. For assessing the proposed approach using testing 274 queries which are the test dataset and examine the right answers by expert domain.

The final results of implementing the proposed approach with the fuzzy logic system enhances the accuracy than without fuzzy logic system (95.7%, 91.6%). In the future work, the “EIWafaa LiFokahaa” corpus will be extended to contain the others Islamic doctrines of jurisprudence then, implementing the proposed approach is applied in the expanded dataset. In addition, it may be apply the neural network as a classifier in the expanded dataset.

REFERENCES


(Fiqh) of worship on the Hanbli’s doctrine], download from http://www.shamela.ws, 1985, last accessed at 5-4-2018
[19] https://studcuy.sharepoint.com/:f:/g/personal/abbeerhassan_pg_cu_educ/Edu/evi0fAWDPApFmFN5TGr78CoBTRBdIsJ6sXd8Il-0mCQ?e=kGAIx7.