mathematics

MDPI

Article

A Short-Patterning of the Texts Attributed to Al Ghazali: A "Twitter Look" at the Problem

Zeev Volkovich

Software Engineering Department, ORT Braude College of Engineering, Karmiel 21982, Israel; vlvolkov@braude.ac.il

Received: 10 October 2020; Accepted: 24 October 2020; Published: 3 November 2020

Abstract: This article presents an novel approach inspired by the modern exploration of short texts' patterning to creations prescribed to the outstanding Islamic jurist, theologian, and mystical thinker Abu Hamid Al Ghazali. We treat the task with the general authorship attribution problematics and employ a Convolutional Neural Network (CNN), intended in combination with a balancing procedure to recognize short, concise templates in manuscripts. The proposed system suggests new attitudes make it possible to investigate medieval Arabic documents from a novel computational perspective. An evaluation of the results on a previously tagged collection of books ascribed to Al Ghazali demonstrates the method's high reliability in recognizing the source authorship. Evaluations of two famous manuscripts, *Mishakat al-Anwa* and *Tahafut al-Falasifa*, questioningly attributed to Al Ghazali or co-authored by him, exhibit a significant difference in their overall stylistic style with one inherently assigned to Al Ghazali. This fact can serve as a substantial formal argument in the long-standing dispute about these manuscripts' authorship. The proposed methodology suggests a new look on the perusal of medieval documents' inner structures and possible authorship from the short-patterning and signal processing perspectives.

Keywords: short-patterning; Al Ghazali authorship; signal processing model; word embedding

1. Introduction and Problem Formulation

This article presents an innovative approach inspired by short-patterning methodologies, aiming to analyze the literary compositions of the outstanding Islamic jurist, theologian, and mystical thinker Abu Hamid Al Ghazali (1058–1111).

1.1. Abu Hamid Al Ghazali

Al Ghazali is one of the most significant Muslim Sufis, whose ideas are prominent and persuasive not only in the Muslim world. As is well acknowledged (see, e.g., [1–4]), he was born at Tus in Persia, where he learned different fields of traditional Islamic religious disciplines. At the age of thirty-three, Al Ghazali was appointed by Nizâm al-Mulk, the Seljuq Empire' powerful vizier, to the Nizâmiyya Madrasa in Baghdad. Afterward, while going through a deep spiritual crisis, Al Ghazali abandoned his excellent career, and in November 1095, he left Baghdad with the excuse of going on a pilgrimage to Mecca. After some time spent in Damascus and Jerusalem, with a visit to Mecca in 1096, Al Ghazali settled in Tus. He spent the rest of his life writing, practicing Sufi, and teaching. In 1106 he went back to the Nizâmiyya Madrasa in Nishapur, where he had been a student, and continued teaching at least till 1110. Afterward, he returned to Ṭus and died the following year. Many modern researchers recognize the significant contributions of Al Ghazali to world theological and philosophical thought.

Al Ghazali had a substantial influence on the development of the Arab–Muslim culture. According to the Hadith predicting the arrival of Islam's renewer once every century, the Arab community

perceived Al Ghazali as the renewer of Islam's fifth century. As an example, the Shafi'i jurist al-Subki claimed, "If there had been a prophet after Muhammad, Al-Ghazali would have been the man."

The Al Ghazali's most meaningful work is *Iḥyāʾ ʿulūm al-dīn* (*The Revival of the Religious Sciences*), primarily considered an outstanding work of Muslim spirituality. The *Ihya* turned out to be the most common Islamic text after the Holy Quran and the Hadith. This book, written in Arabic, is unquestionably essential to individual religious practice and comprehended as one of the greatest works and a timeless outline of the pious Muslim's way to God. Moreover, this extraordinary treatise's outstanding achievement is to unite orthodox Sunni theology and Sufi mysticism together in a valuable, understandable fashion to guide every aspect of Muslim life and death (see, e.g., [1,5,6]).

Al Ghazali's creativity has been the subject of numerous studies and reviews in various Islamic practical aspects and for humanity at large. Many works are attributed to Al Ghazali, occasionally appearing with different titles in different manuscripts (see, e.g., [1,5,6]). This topic is still being explored. The methods applied in these exciting research studies are mainly based on the stylistic and thematical analysis. They involve an in-depth evaluation of the religious and theological views expressed in the works and cross-citation breakdowns. In this regard, it is essential to mention a prominent Scottish Orientalist, historian, academic, and Anglican priest, William Montgomery Watt (1909–2006). His assessments of the authenticity of works attributed to Al Ghazali are considered the most important in this field.

We do not review the explanations devoted to the issue because our purpose is to approach this problem from the formal, mathematical standpoint of modern deep learning methods applied to individual writing style modeling. Hopefully, such a practice can be further merged with the traditional methodologies to combine both attitudes' advantages.

1.2. Authorship Attribution

An individual writing style outwardly expresses an author's perception of reality. It is a personification of the general writing process, composed of many inexact and connecting phases commonly identified as pre-writing, drafting and writing, sharing and responding, revising and editing, and publishing [7]. Therefore, recognizing an individual writing style can be considered to uncover the style's templates, expressed through authorship attributes.

Following this general perception, we consider the assignment of texts prescribed to Al Ghazali from the overall authorship attribution problematics perspective. This field aims to recognize the author of a particular document in question from an analysis of materials with known authorship. A survey of methods applied in this area is given, for instance, in [8]. Such approaches are mainly used in the literature to identify the authorship of novels, plays, or poems with controversial origins.

There are two main kinds of methods in the author's verification problem: intrinsic and extrinsic. The intrinsic methods work merely with the provided texts (one with acknowledged authorship and one undergoing inspection) and form a one-class classification problem. Conversely, extrinsic verification techniques draw the non-target set and create a group of external documents. Extrinsic methods adapt the verification task to a binary classification problem. The most recognized and feasible extrinsic verification approach is the Impostors' method [9].

Supposing we deal with medieval literature, we have to consider the peculiar properties inherent in this type of literary creativity. One of these features is built-in text inhomogeneity caused by multiple frequently unspecified citations of other authors and sources. The manner of expression depends on the target audience and the topic of the text. It may contain many quotations and borrowings; thus, such texts' writing patterns are unstable and vary within the document. Simultaneously, the original style is kept in short patterns inherent to the author (i.e., "The devil is in the details."). The classical procedures for the authorship determination are less accurate when analyzing such short text prototypes.

A similar situation appears in the modern age, where people interconnect through relatively short messages such as tweets. Twitter is a social networking site launched in 2006 to distribute short posts of a maximum of 140 characters, named tweets. The requirement to briefly convey messages as short

tweets has spawned a new literary genre, attracting keen attention from various standpoints. Different applications for the analysis of short texts have been recently proposed to authorize and recognize malicious bots, chat conversations, short message service (SMS) messages, Twitter posts, Facebook status updates. Such studies commonly consist of analyzing short word patterns and should give rise to new methodologies for authorization of a very stylistically heterogeneous textual material. It seems to be very natural to adopt the analytical techniques applied in this area to investigate medieval Arabic texts.

1.3. Paper Contribution

Approaches designed to reveal short patterns in texts using deep learning techniques are known in the literature [10–12], primarily dealing with English language content while paying less attention to others, including ancient languages. As an initial step, these methods involve modifications of word embeddings (see, e.g., [12–17]). Chinese, Persian, Arabic, and Hebrew are significantly distinguishable in their own linguistic and semantic structures from English. In this paper, a preprocessing technique is combined with a word embedding technique constructed for the Arabic language.

The proposed learning procedure follows the general idea of the Impostors' method [14]. This methodology operates with external documents (impostors) and constructs a set of resemblances. In our case, we apply a modified version of the approach. The first basic impostor is composed by the earlier mentioned manuscript, *Iḥyāʾ ʿulūm al-dīn*. The alternative set (the second impostor) includes several books recognized as the Pseudo-Ghazali's ones (composed in imitation of his writing style). Note that these reproductions could be stylistically nonhomogeneous but provided in different ways. Thus, a one-class classification problem is transformed into a binary classification task. Modifying the mentioned deep learning techniques serves as a classifier, trained on the impostors' collection to recognize questionable authorship.

Another difficulty is imbalances in the training material. The size of the fundamental *Iḥyāʾ ʿulūm al-dīn* is expected to be much larger (by a factor of nine) than the total length of the texts from the alternative class. An enriching methodology for alternative balanced groups is constructed and applied to overcome this obstacle in the proposed approach.

The numerical experiments faithfully classify the tested material, previously tagged as written and not written by Al Ghazali, almost entirely corresponding to the accepted perspective. From our standpoint, these results look impressive since they are obtained using a technique utterly different from the accepted ones in this area and entirely based on a formal justification. This study leads to an innovative (signal processing) standpoint on the perusal of ancient documents' inner structures and possible authorship.

At the same time, two known literary compositions, *Mishakat al-Anwa and Tahafut al-Falasifa*, become so close in the sense of their short text features to the so-called Pseudo-Ghazali texts that they are recognized not as written by but only attributed to Al Ghazali. The manuscript *Tahafut al-Falasifa* is commonly acknowledged as Al Ghazali wrote and a student of the Asharite school of Islamic theology. The proposed method recognizes that a prominent part of the text (more than 80%) is written in style substantially differing from that recognized as Al Ghazali's own. Regarding the second manuscript, *Mishakat al-Anwa*, the same conclusion is reached. Note that such an inference is previously made only regarding the final part of the text by Watt [18]. In some ways, these outcomes both confirm and contradict commonly accepted judgments and, of course, have to be compared with future outcomes.

The main contributions of this paper are as follows:

- Adapting a Convolutional Neural Network (CNN) model intended to recognize anonymous authorship using short text patterns;
- Performing detailed analysis of the authorship of creations attributed to Al Ghazali, confirming the reliability of the suggested model;

- Discovering a short pattern structure in two famous works, *Mishakat al-Anwa* and *Tahafut al-Falasifa*, indicating that they very likely to belong to the Pseudo-Ghazali category (merely attributed to Al Ghazali);
- Suggesting a new signal-like text representation to study stylistic text characteristics from the signal processing standpoint.

The rest of the paper is organized as follows. Section 2 states the formal model. In Section 3, the provided numerical experiments are described. Section 4 is devoted to the conclusion.

2. Proposed Method

2.1. Arab Words Embedding

One prevalent text mining method is the bag-of-words technique, presenting the text as vectors of terms' occurrences. This methodology does not preserve semantic information because it ignores the words' orderliness and joint appearances. Moreover, constructed in this way, representations are usually very sparse and suggest additional smoothing techniques. Deep learning embedding systems arrange more exact procedures, providing words' real-value vector representations such that adjacent patterns correspond to words with comparable sense.

Embedding words into a linear space is a trendy modern approach that exhibits semantic and syntactic text properties, implementing the general Distributional Hypothesis (see, e.g., [19,20]), asserting similar meanings of terms appearing in comparative contexts. The work in [21] suggests, based on this principle, the famous Word2vec model of word embedding in real Euclidian space in two fashions: the Continuous Bag-of-Words (CBOW) model and the Skip-gram model. The key idea is to attach a term's feature not to a sole coordinate but an entire compressed vector. Thus, a particular text is translated into a prototype with semantically accomplished columns. Compared with the earlier mentioned bag-of-words procedure, this natural language processing method preserves semantic and syntactic information. These are the most popular embedding methods:

- Word2Vec (by Google) [21];
- GloVe (by Stanford) [22];
- FastText (by Facebook) [23];
- ELMo (AllenNLP's) [24].

CBOW strives to estimate the chance of a word occurrence, using a context such as a solitary word or a words' sequence. Conversely, Skip-gram aims to evaluate the context of a word. Both methods follow the same network topology, yet from opposing directions. The desired representation minimizes, roughly speaking, the distortion between the actual and the predicted matter on a large text corpus. GloVe uses the total word–word co-occurrences estimated on a corpus to reveal a meaningful representation of the word vector space. In comparison with Word2vec, this representation is optimized to approximate the neighboring likelihood logarithm by the words co-occurrences' inner products.

FastText is a modification of the Word2vec model where each word is embodied using characters' n-grams. It is more beneficial, to sum up, the sense of short terms and, likewise, suffixes and prefixes. ELMo is a deep word representation that exhibits upper word features such as syntax and semantics and their evolution across linguistic contexts. A deep bidirectional network qualified on a considerable corpus provides the embedding vectors.

At least 400 million people in about 60 countries consider Arabic as their native language, and about 250 million consider it their second. Arabic is the fifth most spoken language in the world. There are 28 letters in the Arabic alphabet, using only lowercase written letters. Letters sometimes join adjacent ones on both sides or only on the right, thereby creating the Arabic script form named a ligature. A character may appear in up to four different forms, contingent on its location in a word. Arabic is a more complicated morphology than many other languages such as English, French, German, or Russian, but is, to some extent, similar to Hebrew.

One of the most functional Arabic word embedding models is AraVec [25], which offers a pre-trained, word embedding, open-source platform with efficient word embedding models, developed in the general framework of the Word2vec model.

In this way, each term in a pre-trained Arabic word portrayal is substituted with its non-sparse d-dimensional vector representation in the Euclidian space and is trained on modern resources such as Wikipedia or Twitter collections. However, a term from a medieval document may not occur in such texts and may not be captured by the embedding source, thereby excluding it. For instance, these could be words not used in modern language, proper names, or words borrowed from other languages such as Urdu and Persian. These kinds of terms are omitted from our considerations. Considering that the analyzed texts are closer to Modern Literary Arabic than the non-formal Twitter language, we employ in our experiments a 300-dimensional representation trained on the Wikipedia corpus.

2.2. Convolutional Neural Networks

As the tool is intended to discover short patterns in a text, the following neural network architecture is applied. The suggested structure is a CNN, created in the spirit of [10–12], and having as an input, a sequence of matrixes resulted from a word embedding. Let us consider a document

$$D = \{w_1, w_2, \ldots, w_n\}$$

composed from the words w_i, i = 1,..,n and attained from a vocabulary of terms V.

The next ingredient is a convolutional component with the following parameters:

- l—the length of the training sequences;
- l_0—the data batch size;
- $H = \{h_k,\ k = 1, .., t\}$, the sizes of the $1D$ convolution kernels;
- s—the stride size;
- Q—the number of feature maps (kernels having sizes $\{h_i,\ i = 1, .., t\}$);
- N_0—, the number of neurons in the last fully connected layer;
- Emb—an embedding method into $R^{|V|\times d}$ (where the columns correspond to the separate words).

Firstly, we split D into

$$m = \left[\frac{|D|}{l}\right]$$

sequential disjoint parts

$$L_i = \left\{w_{(i-1)*l+1}, \ldots, w_{i*l}\right\},\ i = 1, \ldots, m$$

such that each of them is successively divided into m_0 chunks:

$$m_0 = \left[\frac{l}{l_0}\right].$$

Then, we construct m matrices having the order $d \times l$:

$$G_i = Emb(L_i),\ i = 1, \ldots, m.$$

Here, as before, the matrix columns correspond to individual words.

In the next step, a convolution filter (convolution with a kernel belonging to $h \in H$) is applied to the pieces, obtained via shifting the initial h words in a chunk with an increment equal to the stride size s:

$$Gp_i^{(j)} = \left\{w_{(i-1)*l_0+(j-1)*s}, \ldots, w_{(i-1)*l_0+(j-1)*s+h}\right\},\ i = 1, \ldots, m*m_0,\ j = 1, \ldots, [l_0/s].$$

This yields

$$O_{i,q,k}^{(j)} = F_{q,k} * Gp_i^{(j)}$$

where $F_{q,k}$, $k = 1, \ldots, t$, $q = 1, .., Q$ are the filter matrices of the order $d \times k$. The obtained vectors $O_i^{(j)}$ go through the relu activation function g, applied to create representations in the space $R^{(l_0 - h_k + 1)}$:

$$f_{i,q,k}^{(j)} = g\left(O_{i,q,k}^{(j)}\right)$$

The next step is a max-pooling:

$$z_{i,k}^{(j)} = \max_q f_{i,q,k}^{(j)}, q = 1, .., Q$$

The outcome of this operation is to emphasize the most relevant data across a window. The obtained results are concatenated from the beginning over j and secondly over k, aiming to pass via a fully connected level with $\boldsymbol{N_0}$ components controlling a softmax output layer. As mentioned earlier, the model is designed in the spirit of [10–12]. The main differences are that the network works with words instead of characters' n-grams and uses the relu activation function.

2.3. Handling of Imbalanced Training Data

The amounts of the data located in the two groups are expected to be significantly different due to the immense volume of the basic Al Ghazali's *Iḥyāʾ ʿulūm al-dīn* manuscript forming the authentic main class. Thus, we meet here a typical instance of the imbalanced classification. Such a situation stands, as is well known, as a bias to the majority group, possibly ignoring the minority class overall. This paper builds the following simple procedure, intended to balance the data together with appropriate augmentation. Undersampling of the majority class and oversampling of the minority class are combined before the training. The aim is to balance the training classes involved in the learning procedure.

More precisely, let us suppose that we have two datasets, D_1 and D_2, with $|D_1| > |D_2|$.

Balancing routine (D_1, D_2, F_1, F)
Input parameters:

- D_1 and D_2—the datasets under consideration;
- F_1—the undersampling rate;
- F—the multiplying rate.

Procedure:

- Undersample a sample S_1 from D_1 with the undersampling rate $F_1 * |D_1|$;
- $F * F_1$ times replicate D_2 and get S_2;
- Return S_1 and S_2.

We suggest that F and F_1 are such that $|D_1| > |F * F_1 * |D_1|$. This procedure is anticipated to resolve two problems simultaneously: firstly, to equalize the original set sizes, and then expand them, attempting to stabilize the training process.

2.4. Preprocessing

Preprocessing is a crucial procedure of each NLP (Natural Language Processing) task, especially for Arabic text handling. Such a phase significantly influences the results and has to be matched to the employed embedding method. We operate with the earlier-mentioned AraVec methodology of text preprocessing, which acts in the preprocessing step as follows:

- Remove all punctuation marks, English letters, special characters, and digits;

- Remove tashkeel;
- Remove longation;
- Normalization:
  - Replace (‘ﻭﻭ’, ‘ﻭ’), (‘ﻳﻲ’, ‘ﻱ’), and (‘ﺁ’, ‘ﺍ’);
  - Remove diacritics.

2.5. Procedure

The method is applied to three collections (Cl_i, i = 0,1,2), where Cl_0 includes texts that are unquestionably recognized as written by Al Ghazali. Cl_1 is a collection of books attributed to Al Ghazali but not written by him, and Cl_2 is a tested collection, including at least two anchor books: Al Ghazali definitely wrote the first, and the second is attributed to him. The presence of these anchor items allows interpreting of the obtained clusters of the tested documents. Thus, a cluster containing the first anchor is understood as a collection of the authentic Al Ghazali texts, and the second cluster as-fabricated ones. Algorithm 1 describes the training's overall procedure with the subsequent recognizing the authorship of the tested texts.

Algorithm 1. The proposed procedure's pseudocode.

1. *(Cl_i, I = 0, 1, 2) = Preprocessing step(Cl_i, I = 0, 1, 2)*
2. *(SLi, i = 0, 1, 2) = Dividing (Cli, I = 0, 1, 2) into sequential disjointed chunks of length l_0;*
3. *(Seti, i = 0, 1, 2) = Emb(SLi, i = 0, 1, 2) (Embbeding step)*
4. *Iter = 0 (Initialization of the current iteration counter)*
5. *While(Iter < Niter) do:*
   a. *S0, S1 = Balancing Routine(Set0, Set1, F1, F)*
   b. *Net = Network_training(S0, S1, Parameters) (Traning of the network)*
   c. *IF(Net('accuracy')) < Accuracy threshold*
      then: continue
   d. *Iter = Iter+1*
   e. *For D in Cl_2 do:*
      Labels = Net(Set2(D)) (Avaraged labels of the batchs)
      M(Iter, D) = mean(Labels)
6. *Considering the attained matrix M as a matrix of multivariate data:*
   Each row matches an observation, and each column matches a variable.
7. *Perform a partition of the variables into 2 clusters using the K-Means algorithm:*
   [labels, centers] = K-Means(M,2), where:
   - *labels are assignments of the variables to the clusters*
   - *centers are the centroids of the clusters*
8. *Partition is evaluated using the silhouette method: S= silhouette(labels, M)*
9. *Decision step*
   - *If S < Silhouette threshold or the anchors belong to the same cluster*
     then:
     Documents D∈ Cl_2 are not classified
     stop
   - *otherwise:*
     Documents D∈ Cl_2 are classified according to the anchors' assignment

A few remarks should be made about the algorithm. The process starting at *5.* is performed ***Niter*** times. Item *5a* corresponds to a balancing procedure, providing well-adjusted datasets ***S0*** and ***S1*** using the underlying material $\boldsymbol{D_1}$, $\boldsymbol{D_2}$, to train the proposed neural network in *5b*. The next step, *5c*, checks if the learned result achieves the desired accuracy. If the randomly chosen subset ***S1*** does not provide

the necessary separation, the procedure is repeated with another pair of ***S0, S1***. The loop located in *5.e* classifies all documents from $\boldsymbol{Cl_2}$ via the current iteration network *Net*.

The authentic Al Ghazali class is labeled as *0*, and the alternative is *1*. Thus, ***M(Iter, D)*** represents, for each document from ***D***, the fraction of its parts attained in ***5a*** and recognized as Pseudo-Ghazali. The columns of the matrix *M* are separated in *7.* into 2 clusters using the *K*-means method. Suppose the obtained partition is significant (silhouette is above of Silhouette threshold). In that case, if the anchors' elements are positioned in different groups, then the documents $\boldsymbol{D \in Cl_2}$ are classified according to the anchors' location.

The balancing routine's outcomes strongly depended on the underlying features of the data drawn from the majority class. However, as mentioned earlier, the source class is sufficiently heterogeneous from the stylistic standpoint. Thus, the results are inherently biased toward the primary class due to reproducing its part in the learning process. The scheme utilized here embodies just one possible way to neutralize the bias, using the anchors to recognize appropriate groups.

3. Numerical Study

3.1. Material

In a deep learning model framework, the training material and tested materials consist of texts attributed to Al Ghazali.

1. The source collection (Cl_0) contains text from Al Ghazali's most significant work, *Iḥyāʾ ʿulūm al-dīn*, downloaded from the site http://ghazali.org/ihya/ihya.htm as a collection of 41 files with a total size of 8.5 MB;
2. The alternative collection (Cl_1) include the following texts, with a total size of 1.0 MB:
    1. Al-Ajwiba al-Ghazzāliyya fi 'l-masāʾil al-ukhrawiyya*;
    2. Al-Durra Al-Fakhira*;
    3. Al-Risala al-Ladunniyya*;
    4. Ayyuhal Walad*;
    5. Khulasa al-Tasanif fi al-Tasawwuf*;
    6. Ma'arij al-quds fi madarij ma'rifat al-nafs*;
    7. Miʿraj al-salikin*;
    8. Risalat al-Tayr*;
    9. Sirr al-ʿālamayn wa-kashf mā fī al-dārayn*.
3. The test collection is composed of the following:
    I. Texts agreed upon as written by Al Ghazali;
        1. Al-Mankhul min Taliqat al-Usul* (an anchor);
        2. Al Mustasfa min ilm al-Usul*;
        3. Fada'ih al-Batiniyya wa Fada'il al-Mustazhiriyy*;
        4. Faysal at-Tafriqa Bayna al-Islam wa al-Zandaqa*;
        5. Kitab al-iqtisad fi al-i'tiqad*;
        6. Kitab Iljam Al- Awamm an Ilm Al-Kalam*;
        7. Tahafut al-Falasifa.
    II. Texts agreed upon as not written by Al Ghazali (Pseudo-Ghazali):
        8. Ahliyi al-Madnun bihi ala ghayri*;
        9. Kimiya-yi Sa'ādat* (an anchor);
    III. A book with questionable authorship: *Mishakat al-Anwar* (*Niche of Lights*) (http://ghazali.org/books/mishkat-al-anwar.doc).

Although files containing texts attributed to Al Ghazali are relatively widespread on the Internet, their discovery, collection, cleaning, and validation are painstaking tasks, requiring extraordinary effort and an in-built understanding of the material. The documents marked with an asterisk (*) were prepared and tagged by Ph.D. student Kamal Gasimov, with the assistance and guidance of his

supervisor, Professor of Islamic Studies, Alexander D. Knysh, from the Department of Near Eastern Studies at the University of Michigan, USA. Most of the remaining texts were processed by M.Sc. students of the Software Engineering Programming Department of ORT Braude College, Karmiel, Israel Sami Salami, and Modestos Heib. I would like to express my deep gratitude to all those who contributed to this article's preparation.

3.2. Experiment Setup

The procedure is implemented in Python 3.7.6, using the 2.3.1 version of Keras. The parameters involved in the exhibited research, together with their values in the provided experiments, are as follows:

- *Niter*–the number of iterations = 20;
- Cl_0–the source collection (labeled as 0);
- Cl_1–the alternative collection (labeled as 1);
- Cl_2–the tested collection;
- l–the length of the training sequences = 128;
- F_1–the undersampling rate = 2;
- F–the multiplying rate = 3;
- *Emb*–the words embedding method (the AraVec CBOW 300 dimensional representation, trained on the Wikipedia corpus);
- Accuracy threshold = 0.96;
- Silhouette threshold = 0.75;
- The network (Net) parameters:
  - ✓ T–the number of kernels (3);
  - ✓ H–the sizes of the 1D convolution kernels (3,6,12);
  - ✓ s–the stride size (1);
  - ✓ Q–the number of feature maps (filters) = 500;
  - ✓ N_0–the number of neurons in the fully conned layer = 512;
  - ✓ l_0–the data batch size = 50;
  - ✓ Pooling size = 500;
  - ✓ Learning rate = 0.01;
  - ✓ Momentum = 0.9;
  - ✓ Decay = 1;
  - ✓ Dropout rate = 0.5;
  - ✓ Number of epoch = 10;
  - ✓ Activation function = relu;
  - ✓ Loss = Categorical cross-entropy;
  - ✓ Cross-validation split = 0.25;
  - ✓ Output size = 2;
  - ✓ Optimizer = Adam;
  - ✓ Metric = Accuracy.

3.3. Results

Several numerical experiments are performed, in which the values of the critical parameters l and l_0 vary as 64, 120, 128, and 30, 40, and 50. In general, these trials provide similar results. However, the best outcomes are achieved for $[l/l_0] = 2$. This article's limited scope reports only the most representative results, for $l = 128$ and $l_0 = 50$. In this matter, Cl_0 (the source collection) includes 7619 chunks, with an approximate size of 8.5 MB. Correspondingly, the alternative class contains 819 pieces using about 1.0 MB. As such, the imbalance ratio (IR) is about 8.5.

The training data consists of about 17,700 units, with about 10,000 training, 3300 validation, and 4400 testing samples in each iteration. Following the procedure described earlier, we get training sets that are almost identical in terms of their sizes at each step. I.e., the minor class is multiplied six times. Experiments have shown that this is probably the minimum suitable value. The learning process repeatedly does not converge or exceed the 0.75 validation accuracy for smaller values. Moreover, this characteristic significantly increases in the final stages of training.

Table 1 exhibits the characteristics of the alternative collection.

Table 1. Characteristics of the alternative collection.

	Number of Chunks	Total Size
1	26	32.2
2	95	108.9
3	10	14.8
4	33	39.6
5	49	53.9
6	269	365.4
7	131	169.5
8	7	9.2
9	174	28.5

Documents 6 and 7 dominate in this class.

Based on the Parula color map in the "scaled rows" fashion, a heat map demonstrates the experiments in Figure 1. The document numbers are mapped on the horizontal axis, while the vertical axis represents the experiment number. As can be seen, two text groups are divided by brightness and color into two parts: 1–6 and 7–10. The corresponding cluster procedure separates these sets with the silhouette value of 0.8836.

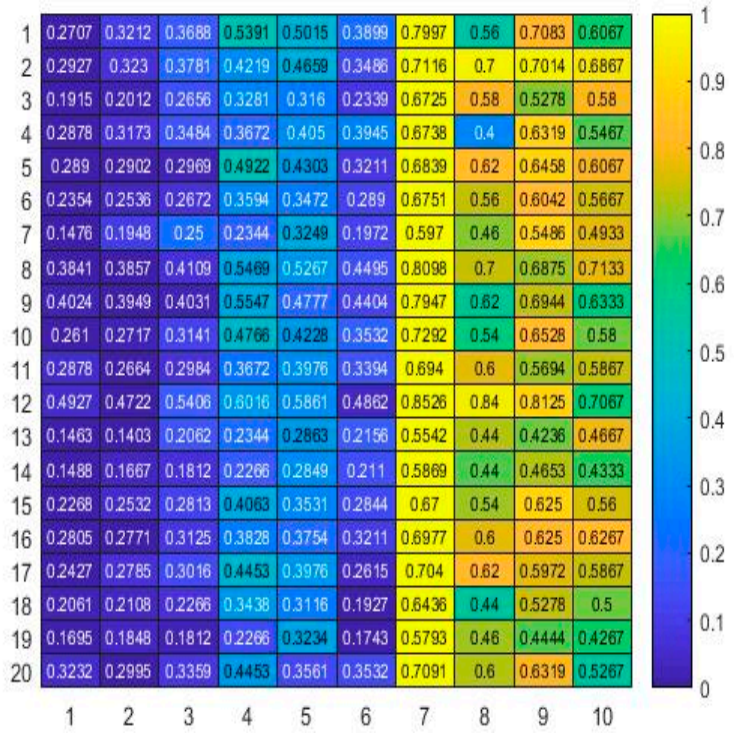

Figure 1. A heat map of the mean values attained with the experiments.

The same situation appears in the error bar charts given in Figure 2. Recall that this graphical representation displays the mean with the variability, specifying by the bars the uncertainty in a measurement. In our case, it embodies one standard deviation. The dotted lines represent the average

cluster centroids (the clusters' averages) $y = 0.3259$ and $y = 0.6090$, respectively. The central line corresponds to the line separating the clusters $y_0 = 0.4674$.

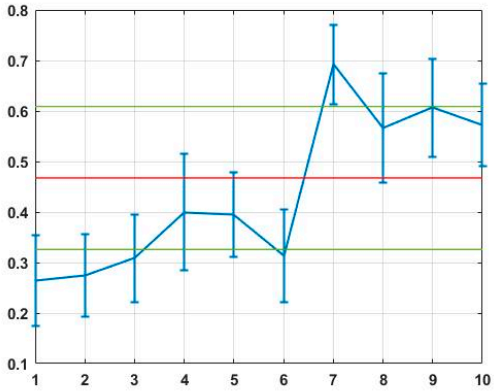

Figure 2. An error bar plot of the mean values attained within the experiments, with one standard deviation.

The partition mentioned above is likewise clearly comprehended here. Let us ascertain this result from the standpoints of the structure of the considered texts. First of all, note that the procedure unquestionably identifies the first six books, known as authorized by Al Ghazali. The procedure tags two books (numbers eight and nine) as Pseudo-Ghazali, perfectly matching the inherent labeling. The two remaining books' classification is of the most significant interest and novelty:

Tahafut al-Falasifa (*The Incoherence of the Philosophers*, seventh in the list of tested manuscripts).

According to the common standpoint, this milestone opus was created by Al Ghazali, together with a student of the Asharite school of Islamic theology. The book criticizes some positions of Greek and other earlier Muslim theorists, mostly those of Ibn Sina (Avicenna) and Al-Farabi (Alpharabius). The manuscript is reputedly an exceptionally successful creation and a landmark in Islamic philosophy.

We explore this topic using additional text representations highlighted by our model. As mentioned before, the procedure divides texts into successive equal-length pieces with the size $l = 128$. According to the predicted classification, each of them is split into batches with the length $l_0 = 50$, tagged as 0 or *1*. In this way, each document D is embodied as a signal, having the length $m = [|D|/l]$, taking $[l/l_0] + 1$ possible values, signifying the pieces tags' mean values. An example of such a signal representation is given in Figure 3. The X-axis represents a piece's sequential number, and the Y-axis shows the average scores of the pieces, which could be 0, 0.5, or 1 in the considered case. Here, $m = 397$; that is, the tested document is divided into 397 pieces, having approximately the same size of 1.2 K. Thus, the numbering on the X-axis is from 1 to 397.

Even here, it can be seen that the Pseudo-Ghazali (the score is above the cluster separation line $y_0 = 0.4674$) covers a more meaningfully significant part of the manuscript.

The overall conclusion has to be based not just on a simple random sample but on the whole assembly of all 20 simulated samples. To do it, we average such curves (signals) obtained in these 20 iterations and consider the resulting sequence, as seen in Figure 4.

The values derived from the averaged series are marked in blue. The red line is the result of the moving average smoothing with lag, equaling 7. This outline characterizes the style's overall behavior, demonstrating that most segments strive to fit the "1" Pseudo-Ghazali style. The observation is also confirmed by histograms generated for the original signal and its smoothed version (see Figure 5).

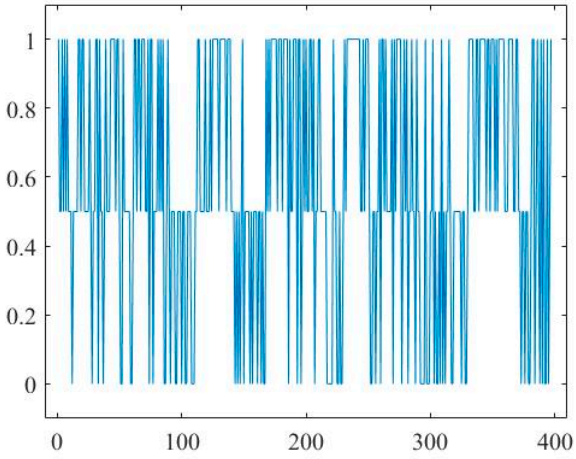

Figure 3. An example of a digital representation of *Tahafut al-Falasifa* chunks.

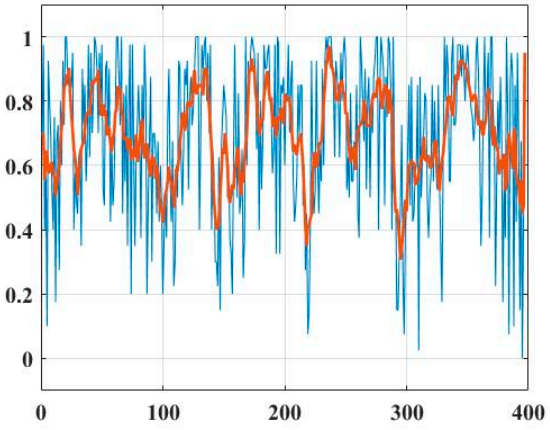

Figure 4. A digitally averaged representation of *Tahafut al-Falasifa* chunks.

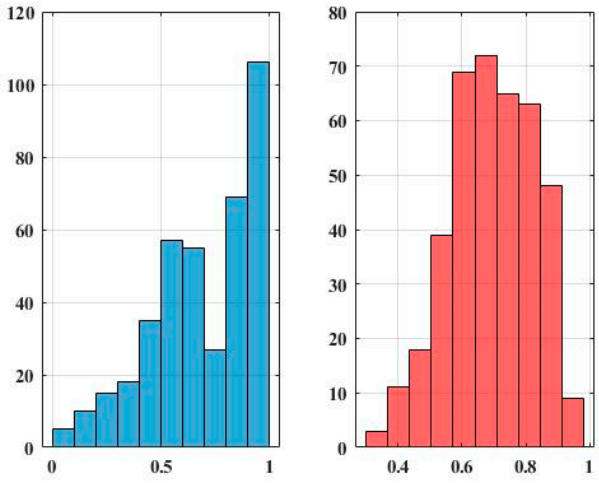

Figure 5. Histograms of the chunked scores of *Tahafut al-Falasifa* for the averaged profile (the left panel) and its smoothed version (the right panel).

Both distributions have a negative skew, specifying a long left tail, the left asymmetry of a distribution around its mean. About 20% of the data are smaller than 0.5 in the left panel and about 8% in the red one. Thus, it is possible to conclude that the dominant part of the considered manuscript *Tahafut al-Falasifa*, is not written in the inherent Al Ghazali style.

Mishakat al-Anwar (The Niche of Lights, number 10 in the tested manuscripts' list)

The prominent official Al Ghazali internet resource (https://www.ghazali.org) dedicates a subsite (https://www.ghazali.org/site/on-mishkat.htm) to the authorship problem of *Mishakat al-Anwar*. Additionally, for several manuscript versions, the site presents the background information and the six crucial papers [5,18,26–28]. These articles apparently can be treated, with some limitations, as the core discussion material in the problem.

The ongoing debate surrounding Al Ghazali's authorship of this manuscript in numerous scientific forums is much more wide-ranging than this website. It refers to documents not mentioned in the current article. In this long-time dispute, the participants present compelling arguments for and against the alleged authorship, based mainly on linguistic, religious, and philosophical outlooks. An analysis and review of these essential issues are not the present paper's subjects because we focus on formal algorithmic methods designed to evaluate the manuscript's authorship.

As in the previous case in Figure 3, we start from an example of a digital signal representation of pieces, given in Figure 6.

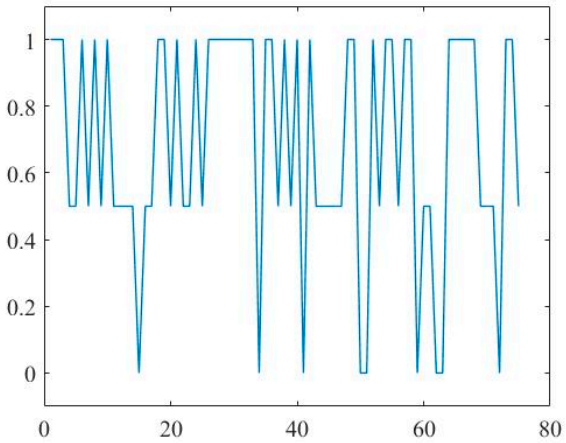

Figure 6. An example of a digital representation of *Mishakat al-Anwar* chunks.

This document is significantly shorter (just 78 pieces) than the one mentioned above; in conclusion, the graph appears to be more sparse. However, the dominance of the scores larger than 0.5 is undoubtedly visible. A chart of the average mean score (blue line) in the trials demonstrates the same tendency in Figure 7. The red line, as previously, corresponds to the moving average smoothing line with lag equaling 7.

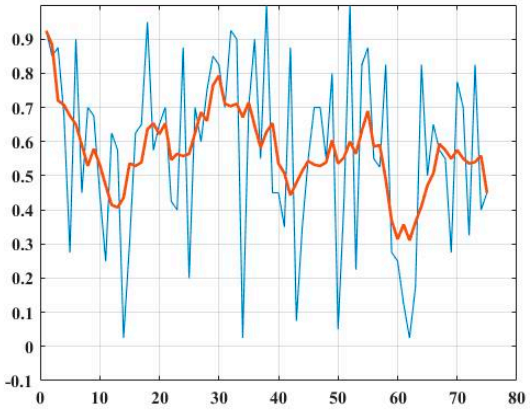

Figure 7. A digitally averaged representation of *Mishakat al-Anwar* chunks.

The resultant histograms also exhibit a left side tail, the left asymmetry of a distribution around its mean (Figure 8).

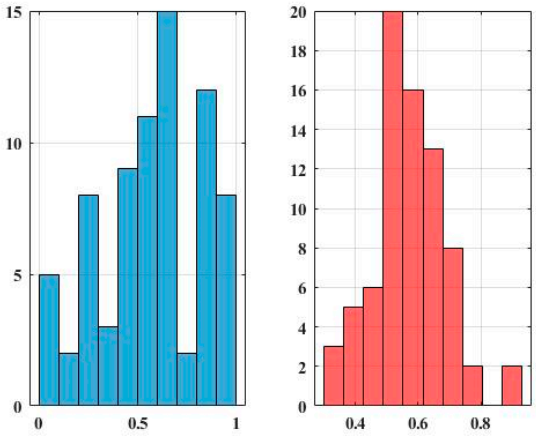

Figure 8. Histograms of the chunk scores of *Mishakat al-Anwar*, found for the averaged profile (the left panel) and its smoothed version (the right panel).

The quantities of the scores lying below 0.5 are 36% and 20%. The general conclusion is that most of the text of *Mishakat al-Anwar* is not composed of the inherent Al Ghazali writing style.

As remarked earlier, we strive to propose a new perspective on the discussed problem. The suggested approach is fundamentally different from those commonly accepted. On the other hand, one study case, in our opinion, is to be debated.

As stated in a paper by Watt [18], "Most of the problems formulated by Gairdner are connected with the last section of the Mishkiit, the detailed interpretation of the Tradition about the Seventy (or Seventy Thousand) Veils (which for convenience I shall call the "Veils-section")." The article [26] of Gairdner is mentioned here. Watt continues, "If the above investigations have not overlooked some crucial point, there is no avoiding the conclusion that the Veils-section of Mishkat al-Anwar is a forgery".

This statement agrees with the results obtained here, where the book's smoothed profile (marked in red in Figure 7) is mostly located above the line $y = 0.5$ in the last part of the chart. As for most of the manuscript, we conclude that this part is not written in the inherent Al Ghazali style. On the one hand, it shows that the obtained results do not contradict the widely accepted opinions. On the other hand, our results generalize them, indicating that the considered book's overall style differs from the inherent one ascribed to Al Ghazali.

4. Conclusions and Discussion

This paper suggests a new approach to the problem of the authenticity of the manuscripts attributed to Al Ghazali. Consideration of the short patterns appearing in the text body makes it possible to unfold a new perspective on the faithfulness and forgery of the studied documents. Combining with a deep learning technique used to analyze tweets, the proposed methodology examines medieval Arabic texts posing high inner inhomogeneity. When the method is applied to the previously tagged text collection, it exhibits reliable results that make it possible to offer a novel text representation in the signal fashion.

The absence of ground truth is an essential but inherent research limitation. Although one generally accepted attitude is confirmed by this study and presented in the last part of the experimental section. Merging the newly proposed method with traditional means can lead to novel inferences for the discussed problem.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Hunt, J. *The Pursuit of Learning in the Islamic World, 610–2003*; McFarland: Jefferson, NC, USA, 2004.
2. Watt, W.M. *Al-Ghazālī, Muslim Jurist, Theologian, and Mystic*; Encyclopædia Britannica, Inc.: Chicago, IL, USA, 2020; Available online: https://www.britannica.com/biography/al-Ghazali (accessed on 1 November 2020).
3. Watt, W.M. *The Faith and Practice of Al-Ghazali*; Revised Edition; Oneworld Publications: London, UK, 1 February 2000.
4. Watt, W.M. *Ghazali/Abu/Hamed/Mohammad, II, III*; Encyclopedia Iranica: New York, NY, USA; pp. 1–12, in print.
5. Wensinck, A.J. Ghazali's Mishkat al-Anwar (Niche of Lights). In *Semietische Studien: Uitde Nalatenschap*; A.W. Sijthoff's Uitgeversmaatschappij, N.V.: Leiden, The Netherlands, 1941.
6. Wensinck, A.J. On the Relation between Ghazali's Cosmology and His Mystiscism. In *Mededeelingen der Koninklijke Akademie van Wetenschappen, Afdeeling Letterkunde 75*; Series A 6; Noord-Hollandsche Uitgevers-Maatschappij: Groningen, The Netherlands, 1993; pp. 183–209.
7. Alred, J.; Brusaw, C.; Oliu, W. *Handbook of Technical Writing*, 9th ed.; St. Martin's Press: New York, NY, USA, 2008.
8. Amelin, K.S.; Granichin, O.N.; Kizhaeva, N.; Volkovich, Z. Patterning of writing style evolution by means of dynamic similarity. *Pattern Recognit.* **2018**, *77*, 45–64. [CrossRef]
9. Koppel, M.; Schler, J.; Argamon, S. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 9–26. [CrossRef]
10. Goldberg, Y. A Primer on Neural Network Models for Natural Language Processing. *arXiv* **2015**, arXiv:1510.00726. [CrossRef]
11. Goldberg, Y.; Hirst, G.; Liu, Y.; Zhang, M. Neural Network Methods for Natural Language Processing. *Synth. Lect. Hum. Lang. Technol.* **2017**, *10*, 1–309. [CrossRef]
12. Prasha, S.; Sebastian, S.; Fabio, G.; Manuel, M.; Paolo, M.; Thamar, S. Convolutional Neural Networks for Authorship Attribution of Short Texts, 6. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Volume 2, pp. 669–674.
13. Aydoğan, M.; Karci, A. Spelling Correction with the Dictionary Method for the Turkish Language Using Word Embeddings. *Eur. J. Sci. Technol.* **2020**, 57–63. [CrossRef]
14. Aydoğan, M.; Karci, A. Turkish Text Classification with Machine Learning and Transfer Learning. In Proceedings of the 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 21–22 September 2019; pp. 1–6.
15. Aydoğan, M.; Karci, A. Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification. *Phys. A Stat. Mech. Its Appl.* **2020**, *541*, 123288. [CrossRef]
16. Ali, F.; El-Sappagh, S.; Islam, S.R.; Ali, A.; Attique, M.; Imran, M.; Kwak, K.S. An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Future Gener. Comput. Syst.* **2021**, *114*, 23–43.
17. Farman, A.; Daehan, K.; Pervez, K.; Shaker, E.; Amjad, A.; Sana, U.; Kye Hyun, K.; Kyung-Sup, K. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowl. Based Syst.* **2019**, *174*, 27–42, ISSN 0950-7051.
18. Watt, W.M. A Forgery in al-Ghazali's Mishkat? *J. Royal Asiat. Soc.* **1949**, 5–22. Available online: https://www.ghazali.org/articles/watt-1949.pdf (accessed on 1 November 2020). [CrossRef]
19. Harris, Z. Distributional structure. *Word* **1954**, *10*, 146–162. [CrossRef]
20. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26, Proceedings of theTwenty-Seventh Conference on Neural Information Processing Systems NIPS, Lake Tahoe, NV, USA, 5–10 December 2013*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2013.
21. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA, 2–4 May 2013.
22. Pennington., J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014.

23. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. FastText.zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
24. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237.
25. Soliman, A.B.; Eissa, K.; El-Beltagy, S.R. AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. In Proceedings of the Third International Conference On Arabic Computational Linguistics, ACLING 2017, Dubai, UAE, 5–6 November 2017; Shaalan, K., El-Beltagy, S.R., Eds.; Elsevier: Amsterdam, The Netherlands, 2017; Volume 117, pp. 256–265.
26. Garidner, W.H.T. Al-Ghazali's Mishkat al-Anwar and the Ghazali Problem. *Der Islam* **1914**, *5*, 121–153.
27. Landolt, H. Ghazali and 'Religionswissenschaft': Some Notes on the Mishkat al-Anwar for Professor Charles, J. Adams. *Asiatische Studien* **1991**, *45*, 1–72.
28. Treiger, A. Monism and Monotheism in al-Ghazali's Mishkat al-Anwar. *J. Quranic Stud.* **2007**, *9*, 1–27.