A Novel Methodology for Topic Identification in Hadith

Sania Aftar[1,*], Luca Gagliardelli[1], Amina El Ganadi[1,2], Federico Ruozzi[1] and Sonia Bergamaschi[1]

[1]University of Modena and Reggio Emilia, Modena, Italy
[2]University of Palermo, Palermo, Italy

Abstract

In this paper, we present our preliminary work on developing a novel neural-based approach named RoBERT2VecTM, aimed at identifying topics within the "Matn" of "Hadith". This approach focuses on semantic analysis, showing potential to outperform current state-of-the-art models. Despite the availability of various models for topic identification, many struggle with multilingual datasets. Furthermore, some models have limitations in discerning deep semantic meanings, not trained for languages such as Arabic. Considering the sensitive nature of Hadith texts, where topics are often complexly interleaved, careful handling is imperative. We anticipate that RoBERT2VecTM will offer substantial improvements in understanding contextual relationships within texts, a crucial aspect for accurately identifying topics in such intricate religious documents.

Keywords

Topic Modeling, Hadith, Neural Topic Model

1. Introduction

In Islamic tradition, the "Hadith", which encompasses the documented sayings and actions of Prophet Muhammad as recorded by his companions, is not only a cornerstone of religious practice but also a subject of considerable linguistic and anthropological interest. These "Ahadith" (plural of Hadith) are integral to Islamic law, acting as a critical secondary source alongside the Qur'an. Their classification into various categories of authenticity—Sahih (authentic), Da'if (weak), Hasan (good), and Mawdu (fabricated)[1, 2] reflects a deep commitment to scholarly rigor and historical accuracy. Each Hadith consists of two fundamental parts: the "Isnad" and the "Matn." The Isnad is the chain of narrators, a lineage of individuals who have transmitted the Hadith through generations, thereby ensuring its authenticity. The Matn, on the other hand, is the core content of the Hadith, encompassing the actual text or message conveyed by Prophet Muhammad. While the Isnad's reliability is pivotal for validating the authenticity of a Hadith,

IRCDL 2024: 20th conference on Information and Research science Connecting to Digital and Library science, February 22–23, 2024, Bressanone, Brixen, Italy

*Corresponding author.
sania.aftar@unimore.it (S. Aftar); luca.gagliardelli@unimore.it (L. Gagliardelli); amina.elganadi@unimore.it (A. El Ganadi); federico.ruozzi@unimore.it (F. Ruozzi); sonia.bergamaschi@unimore.it (S. Bergamaschi)
0000-0001-8151-8941 (S. Aftar); 0000-0001-5977-1078 (L. Gagliardelli); 0000-0002-8196-2628 (A. El Ganadi); 0000-0003-2729-5016 (F. Ruozzi); 0000-0001-8087-6587 (S. Bergamaschi)

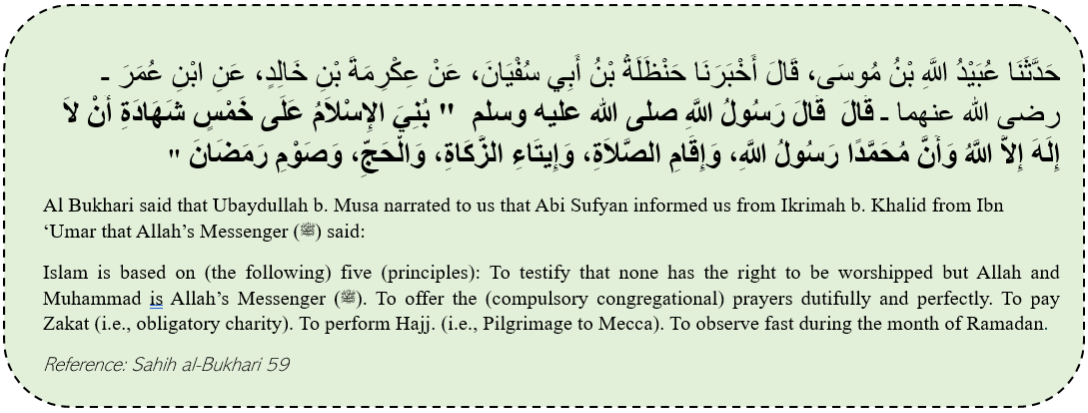

Figure 1: A Hadith presented in Arabic alongside its corresponding English translation. The first part lists the narrators, while the portion in bold represents the Matn.

the Matn provides rich material for theological, linguistic, and anthropological examination.

Traditionally, Islamic scholars have emphasized the scrutiny of the Isnad, sometimes overlooking the profound insights that the Matn can offer. The Matn of a Hadith contains the essence of the Prophet's teachings and actions, offering invaluable perspectives on the linguistic nuances of classical Arabic, as well as the socio-cultural context of early Islamic society. Analyzing the Matn alongside the Isnad allows for a comprehensive understanding of the Hadith, revealing patterns of social interaction, oral traditions, and the transmission of knowledge across generations. This focus has been a defining aspect of the traditional approach to Hadith studies [3]. From a linguistic perspective, the study of Ahadith offers insights into the evolution and usage of classical Arabic, as the language of the Ahadith is often scrutinized for both religious and philological analysis. The precise wording in Ahadith can illuminate subtle nuances of Arabic language and its historical development. Anthropologically, Ahadith provide a window into the social and cultural contexts of early Islamic society. The content of Ahadith, along with the study of their Isnad (the chain of narrators) and Matn, depicted in Figure 1, reveals patterns of social interaction, oral traditions, and the transmission of knowledge across generations. The Isnad, displayed in regular font, represents a lineage of oral transmission, offering clues about social networks and relationships within early Muslim communities. The Matn, highlighted in bold, gives direct insight into the practices, beliefs, and customs of the time. This intricate blend of religious, linguistic, and anthropological elements in the study of Ahadith underscores their multidimensional significance in Islamic scholarship. It highlights the careful balance between preserving religious tenets and understanding the historical and cultural milieu in which these traditions were formed and transmitted.

In recent times, researchers from various fields, including computer science and digital humanities, have shown interest in the subject of Islamic research. Researchers have applied computational techniques such as semantic and named entity resolution approaches with the help of *Natural Language Processing* (NLP) techniques [4] for tackling problems such as identifying the relationship between narrators, resolving ambiguity between narrators reported

by historical Islamic scholars, and answering domain-related queries for Muslims. Additionally, these techniques help in finding a semantic explanation of Matn correspondence to other Hadith, which is also known as content identification [2, 5, 6]. In this context, topic modeling is an ensemble of approaches used to identify patterns and topics in unstructured text. In literature, these approaches are divided into four main categories that fit with different text characteristics, like language and grammar, namely: algebraic, fuzzy, probabilistic, and neural [6]. However, when dealing with non-Latin languages, this task becomes more challenging due to key facts: *(i)* most of the models and libraries are tailored for the English language; *(ii)* the pre-processing is more difficult due to the complexities of word morphology, syntax, and grammar.

In this paper, we propose a novel contextualized model that exploits advanced techniques to address some aspects overlooked by previous approaches. Our work was conducted within the Digital Maktaba project [7, 8] which aims to create an innovative workflow for the automatic extraction of information and metadata from documents in non-Latin scripts (e.g. Arabic). After a comprehensive pre-processing phase, we apply two different embedding techniques: Doc2Vec and Roberta. Doc2Vec provides embeddings for the fixed-size entire document representation, while Roberta generates contextual embeddings for individual words. The embeddings are merged through an autoencoder and subsequently fed into a contextualized topic model. This model takes into account both the outcomes of our proposed embeddings and a bag of words, enabling it to produce semantically categorized topics for Hadith. From a preliminary experimental evaluation, our solution seems to obtain better results concerning the most widely used approaches.

Overall contributions are listed as follows:

- We propose a topic modeling approach named RoBERT2VecTM, designed for the Hadith dataset from Islamic literature, that excels in capturing semantics at both the document and word levels to enhance the accuracy of the generated topics
- RoBERT2VecTM alters the input mode with respect to embeddings that enable it to be suitable for Arabic languages

The subsequent sections of this paper are structured as follows: Section 2 provides a detailed description of our solution. While Section 3 provides related work. Finally, Section 4 delivers conclusions and future directions.

2. RoBERT2VecTM

In this section, we present our semantic-based topic modeling approach for Hadith text. Figure 2 illustrates the workflow diagram of the proposed methodology.

2.1. Hadith Collection

In this study we have selected Hadith books in Unicode format, specifically Sahih Bukhari [1] and Sahih Muslim [2], to facilitate processing. These books are renowned for their authenticity and

[1] http://sunnah.com/muslim
[2] http://www.qaalarasulallah.com/

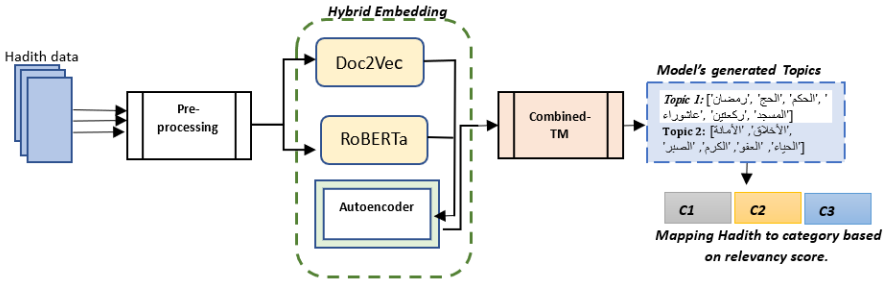

Figure 2: *RoBERT2VecTM architecture.*

i	hadith_id	source	hadith_nc	chapter	Complete Hadith
0	1	Sahih Bukhari	1	Revelation - كتاب بدء الوحي	حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سفيان، قال حدثنا يحيى بن سعيد الأنصاري، قال أخبرني محمد بن إبراهيم التيمي،
1	2	Sahih Bukhari	2	Revelation - كتاب بدء الوحي	حدثنا عبد الله بن يوسف، قال أخبرنا مالك، عن هشام بن عروة، عن أبيه، عن عائشة أم المؤمنين رضي الله عنها أن الحارث بن ه
2	3	Sahih Bukhari	3	Revelation - كتاب بدء الوحي	حدثنا يحيى بن بكير، قال حدثنا الليث، عن عقيل، عن ابن شهاب، عن عروة بن الزبير، عن عائشة أم المؤمنين، أنها قالت أول ما بد

Figure 3: Overview of Hadith Collection

reliability and are sourced from trusted collections. The methodology for our experiment on topic identification involves treating the number of chapters in these books as representative of the number of topics. Accordingly, 50 chapters sharing similar themes will be selected from each book. This method aims to enable a comprehensive and comparative analysis, with the objective of identifying and examining prevalent themes or topics across these texts. Details can be seen in Table 1.

Table 1
Data Collection Sources

Book Name	Total Hadith	Chapters
Sahih Al-Bukhari Book	7370	98
Sahih Muslim Book	7570	57

2.2. Pre-processing

The preparatory stage is crucial for refining the dataset, ensuring it's suitable for task like semantic similarity [9], sentiment analysis [10], and document classification [11] and advanced topic identification. The pre-processing stage provides a clean dataset for the input of our model, after this we got a new single CSV file that contains six columns: *Hadith ID*, *Book name*, *Chapter name*, *Hadith number*, and *Raw text* of Hadith. An initial sample of dataset is depicted by Figure 3. while Figure 4 shows a pre-processing example.

Separating Isnad From Matn. Our study focuses on topic identification, prioritizing text content over narrators. We separate Isnad and Matn using custom regular expressions for targeted text extraction. After the extraction, the rest of the pre-processing tasks are applied to the Matn part only. An example of seperating Isnad from Matn can be illustrated in Step 1 of Figure 4.

Steps	Preprocessing Operations	Samples
1	Separating Sanad and Matan	حَدَّثَنَا عُبَيْدُ اللَّهِ بْنُ مُوسَى، قَالَ أَخْبَرَنَا حَنْظَلَةُ بْنُ أَبِي سُفْيَانَ
		بُنِيَ الإِسْلاَمُ عَلَى خَمْسٍ شَهَادَةِ أَنْ لاَ إِلَهَ إِلاَّ اللَّهُ وَأَنْ مُحَمَّدًا رَسُولُ
2	Text without Diacritics	بني الإسلام على خمس شهادة أن لا إله إلا الله وأن محمدا رسول الله، وإقام الصلاة، وإيتاء الزكاة، والحج، وصوم رمضان
3	Normalization	Initial إأآا with bare (alif) ا Initial ى with (yae) ى ء with ؤ ,ئ (hamza) ة with ه
4	Stop Words Elimination	stop words =[....'ما', 'على', 'هو', 'ان", 'مع', 'ب', 'عن', 'في', 'و', 'على', 'في', 'من']
5	Lemmatization	إسلام على خمس شهاد أن لا إله إلا الله أن محمد رسول الله

Figure 4: Doc2Vec Cluster Visualization in 2D plane

Removal of Diacritics. Removing diacritics, marks that show linguistic traits, simplifies text for more effective topic modeling. An example can be seen in Figure 4–Step 2

Normalization. Normalization improves text data for analysis by standardizing Unicode characters for model compatibility and reducing word repetition to unify word forms, as demonstrated in the Figure 4–Step 3.

Stop words Removal. Stop-word removal enhances topic model precision by focusing on significant words. We employed *Python's nltk.corpus* for this purpose and enhanced accuracy by customizing the stop-word list.Some stopwords are highlighted in Step 4 of Figure 4.

Lemmatization. Lemmatization is the process of reducing different forms of a word to its base form (or lemma) [12]. We use the Farasa lemmatizer[3], recognized as one of the standardized and freely accessible tools in comparison to other lemmatizers [12]. Figure 4–Step 5 shows an example of lemmatization.

2.3. Contextualize Model for Hadith

We intend to employ a hybrid embeddings to a neural-based model, following the pre-processing steps, to effectively identify and categorize content within Arabic Hadith texts.

Hybrid Embedding. After performing previous described pre-processing on our dataset, we obtain a clean version that is then used to generate embeddings. Creating a semantic-based vector representation of the entire document poses a challenge, as traditional techniques such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are limited in providing a semantic numerical representation of the input. To overcome this challenge, we use a hybrid embedding technique to create a semantically enriched vector representation of the Hadith text.

Given the Matn of a Hadith, the hybrid embedding approach combines Doc2Vec [13] and a RoBERTa model [14], fine-tuned for Arabic text, to capture its semantic in two different ways: (*i*) Doc2Vec generates embeddings covering the entire content based on the similarity of the words producing an embedding of a 128-size; where as RoBERTa obtain a more nuanced understanding providing the embeddings of each word in the surrounding context. Given a text, RoBERTa first converts it into tokens, then returns an embedding of 768-size for each of them, resulting in a matrix of (#tokens x 768)-size. To be able to combine the embeddings from both models

[3]farasa-api.qcri.org

they must have the same dimensions, to this end we use Uniform Manifold Approximation and Projection (UMAP) [15] to process the matrix obtained from RoBERTa obtaining a 128-size embedding that preserves the original semantics.

Finally, we use an auto-encoder to analyze and normalize both the embeddings to obtain a unique one, ensuring that all features contribute equally to the learning process. The auto-encoder is initialized with suitable input and bottleneck layer sizes that are calculated using Mean Squared Error (MSE) loss and the Adam optimizer during training. For 200 epochs, the model parameters are updated to learn a compressed representation. The resulting encoded embeddings are then used in our proposed model.

Combined Topic Model. Giving a document, the Combined Topic Model (CTM) [16] determines its topics by combining a Bag-of-Words (BoWs) representation and the embeddings generated with Sentence-BERT(SBERT)[17]. In our solution, we have replaced SBERT with our hybrid embeddings obtained as described in the previous sub-section.

Initially, by treating various chapters as topics, the model produced topics with semantic relationships. Subsequently, we undertook a cleanup process to decrease their quantity by eliminating duplicates and irrelevant terms, including names or words shorter than two characters. After selecting eight categories according to themes, we organized our filtered topics within these categories to provide a clearer understanding of the context. For example, topics concerning to compulsory obligations in Islam were combined into a single one labeled as "Prayer and Worship", while themes regarding moral valueswere merged under "Islamic Ethics and Morality".

Preliminary Results. We performed a preliminary evaluation of our solution, obtaining a higher coherence score [18] than those obtained with other baseline solutions, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) . However, since this is a work in progress, we plan to publish them in the future (see Section 4).

3. Related Work

In this section, we provide related work, particularly that targets topic modeling, Machine Learning (ML), or NLP for Hadith.

Topic modeling. Advancements in topic modeling, particularly in sentiment analysis and text summarization [19], have been driven by neural network models. These neural models enhances the traditional approaches by integrating models with different embeddings for better topic construction [20, 21, 22]. DeepLDA, a novel neural network proposed by Bhat et al. [23] notably reduces computational time in LDA for topic identification, surpassing the performance of traditional LDA.

Numerous research works have included context-based topic modeling techniques. In order to improve model coherency, Zhao et al [21] introduced the Variational Auto-Encoder Topic Model (VAETM), which combines entity and word vector representations. Xie et al. [24] applied a multilingual BERT-enhanced LDA model to study topic trends in multilingual scientific texts. Similar to this, Habbat et al. [25] combined Product of Latent Dirichlet Allocations (ProdLDA) with AraBERT, an Arabic BERT model, achieving more coherent topics than with traditional LDA and other models.

Content-Based Study of Hadith. Al-Kabi et al. [26] proposed research on classifying Hadith by chapter titles in Sahih Al-Bukhari, by applying TF-IDF for document weight assessment. The study pointed out the challenge in semantically categorizing Hadith. Another study by Nohuddin et al. [27] used text mining and cluster analysis to explore word interrelationships in Hadith chapters, focusing on keyword frequency and similarities across chapters. Najid et al. [28] applied Support Vector Machine (SVM), naive Bayes, and k-Nearest Neighbors (k-NN) classifiers to evaluate Malay-translated Hadith based on Isnad. Similarly, Abdelaal et al. [29] uses a different approach to enhance Hadith classification accuracy by utilizing different supervised learning algorithms. Overall, we consider that our proposed strategy for topic identification, that involves semantic based identification of Hadith using hybrid embedding makes it more efficient and accurate.

Narrator Based Study. The Shamela library[4] is useful for Hadith study, provides in-depth chain of narrator analysis but can't automatically differentiate between 'Sahih' and 'Da'ief' Hadith. Mostly studies put significant emphasis on the Isnad however, some of them establish methodologies based on different models to classify Hadith based on narrators [5, 30, 31, 32, 33]. In our study, we focus on the Matn part of Hadith, considering topic modeling and checking the coherency of the proposed solution with alternatives. We only use the narrator as a source of Hadith.

4. Conclusion and Future Work

In this work, we proposed RoBERT2VecTM, a novel topic identification model for the Arabic language that has the potential to outperform current state-of-the-art approaches in the context of Hadith. By using a custom embedding technique that combines token-based and document-based embeddings, in some initial experiments, we found that the model effectively captures the unique linguistic and thematic elements inherent in Islamic texts. This approach might not only enable the identification of key terms but will also provide insights into their broader significance within the Islamic tradition. In the future, we plan to extend this work with a complete experimental evaluation to demonstrate the capabilities of RoBERT2VecTM.

Acknowledgments

This work was conducted within the PNRR project "ITSERR - Italian Strengthening of the ESFRI RI RESILIENCE" (Avviso MUR 3264/2022) funded by EU – NextGenerationEU - Grant No IR0000014.

References

[1] A. Mahmood, H. U. Khan, F. K. Alarfaj, M. Ramzan, M. Ilyas, A multilingual datasets repository of the hadith content, International Journal of Advanced Computer Science and Applications 9 (2018) 165–172.

[4] http://shamela.ws (Accessed on 16 Jul 2023)

[2] M. Mghari, O. Bouras, A. El Hibaoui, Sanadset 650k: Data on hadith narrators, Data in Brief 44 (2022) 108540. URL: https://www.sciencedirect.com/science/article/pii/S2352340922007478. doi:https://doi.org/10.1016/j.dib.2022.108540.
[3] J. Schacht, The origins of Muhammadan jurisprudence, Oxford University Press, 1967.
[4] A. M. Azmi, A. O. Al-Qabbany, A. Hussain, Computational and natural language processing based studies of hadith literature: a survey, Artificial Intelligence Review 52 (2019) 1369–1414.
[5] M. M. A. Najeeb, Towards a deep leaning-based approach for hadith classification, European Journal of Engineering and Technology Research 6 (2021) 9–15.
[6] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, A. Hassan, Topic modeling algorithms and applications: A survey, Information Systems 112 (2023) 102131.
[7] R. Martoglia, S. Bergamaschi, F. Ruozzi, M. Vanzini, L. Sala, R. A. Vigliermo, Knowledge extraction, management and long-term preservation of non-latin cultural heritages - digital maktaba project presentation, in: A. Falcon, S. Ferilli, A. Bardi, S. Marchesin, D. Redavid (Eds.), Proceedings of the 19th The Conference on Information and Research science Connecting to Digital and Library science, IRCDL 2023, Bari, Italy, February 23-24, 2023, volume 3365 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 153–161. URL: https://ceur-ws.org/Vol-3365/short11.pdf.
[8] S. Bergamaschi, S. D. Nardis, R. Martoglia, F. Ruozzi, L. Sala, M. Vanzini, R. A. Vigliermo, Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach, Sensors 22 (2022) 3995. URL: https://doi.org/10.3390/s22113995. doi:10.3390/S22113995.
[9] M. O. Alhawarat, H. Abdeljaber, A. Hilal, Effect of stemming on text similarity for arabic language at sentence level, PeerJ Computer Science 7 (2021) e530.
[10] R. Duwairi, M. El-Orfali, A study of the effects of preprocessing strategies on sentiment analysis for arabic text, Journal of Information Science 40 (2014) 501–513.
[11] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. Al-Qaness, M. Abd Elaziz, A. Dahou, A study of the effects of stemming strategies on arabic document classification, IEEE access 7 (2019) 32664–32671.
[12] H. Mubarak, Build fast and accurate lemmatization for arabic (2017).
[13] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.
[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[15] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).
[16] F. Bianchi, S. Terragni, D. Hovy, Pre-training is a hot topic: Contextualized document embeddings improve topic coherence, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 759–766. URL: https://aclanthology.org/2021.acl-short.96. doi:10.18653/v1/2021.acl-short.96.
[17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks,

arXiv preprint arXiv:1908.10084 (2019).
[18] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 262–272.
[19] A. P. Logan, P. M. LaCasse, B. J. Lunday, Social network analysis of twitter interactions: a directed multilayer network approach, Social Network Analysis and Mining 13 (2023) 65.
[20] H. Larochelle, S. Lauly, A neural autoregressive topic model, Advances in Neural Information Processing Systems 25 (2012).
[21] X. Zhao, D. Wang, Z. Zhao, W. Liu, C. Lu, F. Zhuang, A neural topic model with word vectors and entity vectors for short texts, Information Processing & Management 58 (2021) 102455.
[22] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, A. Candelieri, Octis: Comparing and optimizing topic models is simple!, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021, pp. 263–270.
[23] M. R. Bhat, M. A. Kundroo, T. A. Tarray, B. Agarwal, Deep lda: A new way to topic model, Journal of Information and Optimization Sciences 41 (2020) 823–834.
[24] Q. Xie, X. Zhang, Y. Ding, M. Song, Monolingual and multilingual topic analysis using lda and bert embeddings, Journal of Informetrics 14 (2020) 101055.
[25] N. HABBAT, H. ANOUN, L. HASSOUNI, Arabertopic: A neural topic modeling approach for news extraction from arabic facebook pages using pre-trained bert transformer model, International Journal Of Computing and Digital System 14 (2021).
[26] M. Naji Al-Kabi, G. Kanaan, R. Al-Shalabi, S. I. Al-Sinjilawi, R. S. Al-Mustafa, Al-hadith text classifier, Journal of Applied Sciences 5 (2005) 584–587.
[27] P. Nohuddin, Z. Zainol, K. Chao, M. Tarhamizwan, S. Marzukhi, A. Nordin, Keyword based clustering technique for collections of hadith chapters, International Journal on Islamic Applications in Computer Science And Technologies–IJASAT 4 (2016) 11–18.
[28] M. N. SR, N. Abd Rahman, N. Alias, M. Alias, et al., Comparative study of machine learning approach on malay translated hadith text classification based on sanad, in: MATEC Web of Conferences, EDP Sciences, 2017.
[29] H. M. Abdelaal, B. R. Elemary, H. A. Youness, Classification of hadith according to its content based on supervised learning algorithms, IEEE Access 7 (2019) 152379–152387. doi:10.1109/ACCESS.2019.2948159.
[30] M. M. A. Najeeb, A novel hadith processing approach based on genetic algorithms, IEEE Access 8 (2020) 20233–20244.
[31] M. M. A. Najeeb, Xml database for hadith and narrators, American Journal of Applied Sciences 13 (2016) 55–63.
[32] R. Yotenka, S. K. Dini, A. Fauzan, A. Ahdika, Exploring the relationship between hadith narrators in book of bukhari through spade algorithm, MethodsX 9 (2022) 101850.
[33] A. Ramzy, M. Torki, M. Abdeen, O. Saif, M. ElNainay, A. Alshanqiti, E. Nabil, Hadiths classification using a novel author-based hadith classification dataset (abcd), Big Data and Cognitive Computing 7 (2023) 141.