Sümeyye Akça

# 15 Topic Modelling in the Ottoman Kadi Registers

**Abstract:** The Qāḍī/Qadi/Kadi is the judge who renders decisions in Islamic Shariah law. The registers documenting the court decisions by the Kadis also describe daily events in the life of the Ottoman Empire and are of great importance for Ottoman research. The data recorded in the registers highlighted the problems of social life of the period and provide modern-day researchers and interested people the opportunity to penetrate the details of the cultural life of the Ottoman state. Using computational methods enables a close reading of the records by historians and other social researchers which significantly enhances the value of the registers and leads to improved research outcomes. Natural Language Processing (NLP) methods like topic modelling and clustering have been used to identify topics in large document collections and the relationships between them. This chapter outlines the work undertaken in a project which used topic modelling to analyse the Kadi registers. The resulting modelling identified connections between the descriptive words used to record the subject matter in each record within the register and the connections were used to form a subject clustering list. Topics identified were compared with previously determined subjects and the semantic links between registers revealed. The project goal was to enable an effective approach for ensuring effective searching and information retrieval in the Kadi registers database.

**Keywords:** Qadi registers; Kadi registers; Natural language processing; Ottoman empire – Social life and customs; Topic modelling

## The *Qāḍī*/Qadi/Kadi Registers

The *Qāḍī*/Qadi/Kadi is the magistrate or judge of an Islamic *sharīʿa* court, who also exercises extrajudicial functions. The Kadis recorded their decisions in registers or notebooks which constitute the court records and can be defined as "a book which has the records of any document created as a result of the legal cases heard by *kqadis* (Islamic judge) as well as administrative and legal activities of qadis" (Aydin and Tak 2019). Kadi Registers contain a wide range of information on the life and times in which court decisions were recorded and shed light on social life in the Ottoman Empire. They are one of the most important sources for today's historians (Akça 2005). They were written in Turkish, Arabic, and Persian and maintained in book format. The books, which were maintained and updated from


https://doi.org/10.1515/9783111336435-016

the second half of the 15th century through to the first quarter of the 20th century, are one of the key sources of information on Turkish culture and history, and are also closely linked to Turkish economic and political life. Over the last 50 years many researchers in Turkey have been transcribing the content of the Kadis to the Latin (Roman) alphabet to facilitate studying the wealth of content contained in the registers.

A significant project, the Istanbul Kadi Registers Project, was undertaken by the *İslam Araştırmaları Merkezi*/ Centre for Islamic Studies (İSAM) working with other organisations and resulted in the transcription of a large corpus of 60 notebooks from 1557–1911, with a large body of the content available in an online database, the *İstanbul Kadı Sicilleri*/Istanbul Kadi Registries." In addition to being historical data, Kadı Registers are an extremely important and reliable source for many scientific fields such as law, economics, geography, sociology and psychology" (Istanbul Kadi Registries n.d.). Each book contains records for every unique event, giving clues to the legacy structure of the Empire. Although to date there has been significant effort to transcribe the notebooks, some still await transcription into the Latin alphabet.

Following transcription into the Latin alphabet, a subject descriptor has been assigned by the transcriber to each record in the register. Researchers can search the indexed registers by subject and retrieve relevant content. However, one record may refer to more than one subject area, and in some cases, a particular event may be covered by more than one record. Assigning appropriate subject headings to the content of each entry in the registers through the use of a thesaurus of terms opens up the content to researchers. Creating subject authority files is considered valuable in terms of providing the basis for indexing the content of the notebooks thereby rendering the content discoverable, accessible and available to wider audiences. It is important to define and analyse the content of information resources according to subject authority files to provide more effective results in information retrieval. Subject authority files are important sources in the field of information science. They provide standard terms from a controlled vocabulary and can be used by indexers and users alike to understand how content can be described or sought (Gültekin 2020, 47).

The aim of the project described in this chapter was to automatically assign a topic to each record in the Kadi registers using the topic modelling method and to reveal the distinctive dominant word groups of the topics by examining the common word groups used for each topic. The multiple records pertaining to a specific subject area can provide valuable insights into the structure and content of the Kadi registers. Analysing the descriptive terms associated with each record and examining word frequency allows for the identification of interconnected subject clusters. When users search using specific terms, a structured hierarchy provides

researchers with access to a broader range of topics, including records related to those terms. It is crucial to ensure that the Kadi registers, which hold significant historical value, are effectively used by a diverse audience of users. A comparison was made with topics tagged during manual transcription. The project involved the creation of an authority file that could be used when describing the content of the records contained. The ultimate goal of the project is to facilitate discovery and access of the information contained in the Kadi registers.

## Machine Indexing and Analysis of Historical Documents

Projects in which computer or computational methods have been used on historical data and documents have been carried out with the intention of making the data and information contained in the documents easier to understand and interpret. Grant et al. (2021) used topic modelling to undertake an historical analysis of global policies on refugees within approximately 55,000 pages from the 1970s of typewritten and digitally born documents stored in archives from the UK and US governments and the United Nations High Commissioner for Refugees (UNHCR). Standard optical character recognition (OCR) was used to scan the documents and natural language processing (NLP) topic modelling was adopted as the means of identifying topics within the documents and their relationships. Topic modelling is a text-mining tool used to identify semantic structures in a given body of text by statistically examining the occurrence of particular words. A document typically refers to several topics in unequal proportions. Clusters of similar words are captured using topic modelling. Topic modelling can help organise the plethora of written material produced today which cannot be analysed by human means and contribute to understanding large collections of unstructured text. The study analysing the documents on refugees identified the main topics in each document and investigated the transmission of the topics between organisations and over time to suggest areas for further study. Major themes and varying organisational approaches were identified. Researchers were then able to analyse the roles of the various people and organisations involved and their discourse, as well as the results of resettlement programmes.

In another study, Yang, Torget, and Mihalcea (2011) applied topic modelling to the content of newspapers published in Texas US between 1829 and 2008. The study examined the difficulty historians experience when working with the huge quantity of digital data available for analysis which has been generated from the projects transferring newspaper content to the digital environment. Topic model-

ling was used to cluster topics within the digital texts to identify the most important and potentially interesting topics, and/or to discover unexpected topics through unusual patterns over a given period of time. Researchers can use the clusters to focus on the topics most closely identified with their research. Schöch (2017), used topic modelling to examine French drama texts between the years 1610-1810, examined whether different dramatic genres had distinctive dominant themes and whether sub-genres have their own specific plot patterns. Schöch's study looked at the extent to which clustering and classification methods based on subject scores produced results consistent with traditional genre distinctions. The various studies undertaken have demonstrated that topic modelling is an effective means of analysing historical documents for further research examination.

## Digitisation of Ottoman Manuscripts

Making Ottoman manuscript documents and books readable and accessible by computer is a process that requires serious effort. There is no system that works with a high accuracy rate among existing studies. For various reasons, Ottoman manuscripts seem particularly difficult to read by machine. There are issues with the handwriting and language used and with the condition of the manuscripts themselves. The language of *Osmanlı Türkçesi*/Ottoman Turkish has particular issues being based on Arabic. One of the most significant difficulties is that Ottoman manuscripts were not written on paper with a satisfactory flat surface or plane, which makes scanning of the content challenging.

In the literature, there are studies using different methods to render irregularly written manuscript historical documents machine-readable and accessible. With advanced information technology, it is possible to extract data from documents by using machine learning (ML) algorithms. Existing common techniques based on handwritten text recognition (HTR), combined with deep learning (DL) recurrent neural networks (RNN) have greatly contributed to the ability of machines to scan, recognise, analyse and store the content of complex manuscripts. DL architectures are used to detect line and page structures of documents, and various projects have applied techniques to extract line and page structures with methods such as masking used with convolutional neural networks (CNN), RNNs and graph neural networks (GNN). CNNs and RNNs have different architectures with one feeding data forward and the other back and are commonly used to solve problems involving spatial data, such as images. RNNs are better suited to analysing temporal and sequential data, such as text or videos (Craig and Petersson 2023). Masking refers to skipping missing data during machine processes (Zhu and Chollet 2023).

Deep neural networks and artificial intelligence have significantly accelerated the automatic transcription of historical manuscripts. Various studies have adopted the techniques mentioned to analyse handwritten texts (Andrés et al. 2022; Ares Oliveira 2018; Gao et al. 2019; Gilani et al. 2017; Prasad et al. 2019; Qasim, Mahood, and Shafait 2019; Siddiqui et al. 2019). These studies have demonstrated effective use of artificial intelligence techniques to the analysis of text but clearly also demonstrate the need to use structured automated methods that take the context into account to reduce both layout analysis errors and text recognition errors (Prieto et al. 2023). Existing HTR implementations are not error-free (Andrés et al. 2022).

Software and platforms using HTR and DL technology like Transkribus have been developed and used for various projects. "Transkribus is an AI-powered platform for text recognition, transcription and searching of historical documents – from any place, any time, and in any language" (Read Co-op 2023). However, varying document layouts, types of handwriting or fonts used, and differences due to the person writing the article make the task of effective transcription and analysis challenging even with the use of appropriate software (Lang et al. 2018). In particular, document image understanding is a difficult pattern recognition problem extracting and distinguishing relevant data that requires complex models based on ML. The problem is even more demanding for document images with complex layouts, such as tables. The reading order is often inherently ambiguous and, as a result, the context is often unclear (Prieto et al. 2023). Document layout analysis focuses on identifying and categorizing the regions of interest in the scanned image of a text document and recent developments have potential for improvements in image analysis with geometric and logical layout analysis, and top-down approaches which divide a document into columns and blocks based on white space and geometric information in addition to bottom-up approaches which examine the raw pixel data. Accuracy rates in pattern recognition above average have been found in trials using the hidden Markov model (Motawa, Amin, and Sabourin1997; Onat, Yildiz, and Gündüz 2008). The improvements give impetus to undertaking further studies on the viability of automated means of analysing Ottoman manuscript documents.

Since printed works have a page layout, it seems easier to digitise them through machine-readable means by scanning and use of AI techniques, and make the content more widely accessible. A successful Turkish project has been:

> Wikilala, nicknamed as Google of Ottoman Turkish, is a Turkish digital library of Ottoman Turkish textual materials. Wikilala, which is currently in its beta version, consists of more than 109,000 printed Ottoman Turkish textual materials, including over 45,000 newspapers, 32,000 journals, 4,000 books and 26,000 articles. Wikilala, provides its users with full-text search through its database using Ottoman Turkish alphabet or Turkish alphabet (Wikipedia 2024).

The issues with digitising and transcribing Ottoman Turkish based on Arabic have been mentioned. A *Luggat Osmanlica Türkçe Sözlük*/Luggat Ottoman Turkish Dictionary with Arabic and Persian spelling and Ottoman pronunciations and detailed explanations has been developed and provides support (*Luggat Osmanlica Türkçe Sözlük* n.d.). Another project impacting on the Kadi register project has been the development of an Ottoman transcription tool.

> Akis is carried out in cooperation with DH Lab and VERİM (Data Analytics Research and Application Center). Our aim is to develop recognition technologies that can transcribe Ottoman Turkish handwritten and printed works written in Arabic and Persian script into Latin script. Thus we want to make texts in Ottoman archives and libraries written in Ottoman Turkish more accessible to researchers and general users from different disciplines (*Sabancı Üniversitesi* n.d.).

Various issues encountered in working with digitisation in relation to Ottoman manuscripts were highlighted in the study undertaken by Kirmizialtin and Wrisley (2022). Their study of a digital newspaper collection printed in Arabic-script Ottoman Turkish in the late 19th and early 20th centuries emphasised the difficulties of automated transcription of non-Latin script languages along with difficulties in training HTR models, particularly with writing systems which have experienced change over time. Synthetic data and other data have been used in studies to augment data available for input to ML models. A test of data augmentation on printed works in Ottoman Turkish found that the system could be improved if there was a larger database (Bilgin Tasdemir 2023).

In projects carried out all over the world, documents and records that cannot be read by machine due to line and page formats and other issues continue to be transcribed and made accessible by volunteers. Crowdsourcing is currently the most likely method for extending the digitisation and analysis of manuscripts. Zooniverse hosts projects for people-powered research where volunteers work with researchers to build datasets for analysis. FromThePage is a crowdsourcing platform for archives and libraries where volunteers transcribe, index, and describe historic documents.

## Data Set and Research Methods Used

This section outlines the details of the project involving the Kadi registers which focused on one of the registers, the book of Üsküdar, a district within Istanbul. It covers the years 1561-1563, and had previously been transcribed by the author. The register contains important information about the social life of Üsküdar, with 648

*hüccets*/court records such as purchase and sale transactions, cases about fugitive slaves, and murders (Akça 2015, 9). In the transcribed version of the book, a manual classification had been given for the topic of each court record, with an index consisting of subject, person, event and place information. A topic modelling algorithm was run on the text to machine-classify the topics of the court records in the text and to compare them with previously assigned topic headings.

As previously outlined in this chapter, topic modelling is an unsupervised approach to finding groups of words in a text document. The topics are made up of words that often occur together and share a common theme. Topics with a predefined set of words can be used as phrases to describe the entire document. To improve the identification of the subject of each record, Latent Dirichlet allocation (LDA) was run on the text using Python software. LDA is a type of topic modelling algorithm where each document is considered a collection of topics with each word in the document corresponding to one of the topics.

The study commenced by importing the data and running pre-processing tools on the text. The initial clean function was used to remove the punctuation marks on the text and change uppercase letters to lowercase. In addition, word tokenization was used to demarcate words and symbols with spaces. The stop word removal function was used to remove the most commonly used words in a language. The open-source Gensim library was used to create the corpus and a dictionary. Gensim is an open-source library for unsupervised topic modelling and NLP using modern statistical ML.

To apply topic modelling, the number of topics must be determined. There are many approaches currently available for this task. The approaches generally look at the distributions of LDA such as subject-terms, document-subjects, and calculate the distances between pairs of topics and determine the most appropriate number of topics (Akbulut 2022, 29). Four metrics were used to determine the most appropriate number of topics for the LDA algorithm to be applied on the Kadi register. The metrics included statistical calculations based on topic and document calculations on the text, and each metric was run in R-project over a single code. As a result, the most suitable number of topics for the model to work on the data set was determined to be thirty (Figure 15.1).

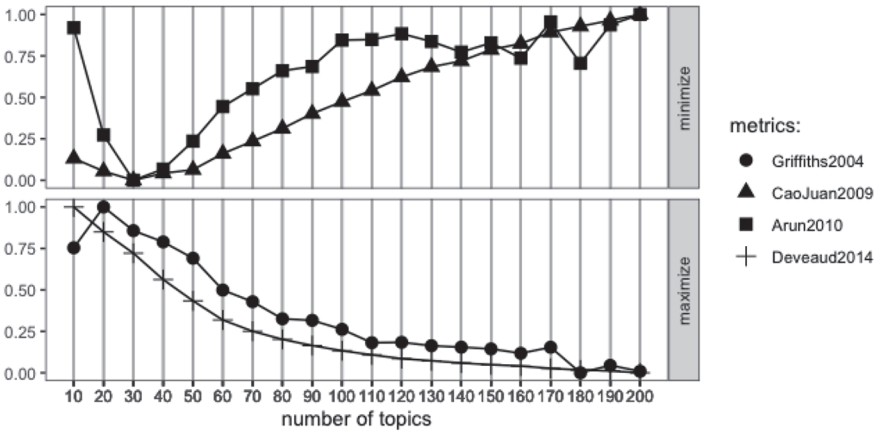


**Figure 15.1:** Four metrics used to determine the optimal number of topics for the LDA model

The topic models were then created and the results analysed. PyLDAvis was used to visualise the results. PyLDAvis is designed to help users interpret topics in a topic model as appropriate to a collection of text data. It provides an interactive web-based visualisation with information from the LDA topic model. Finally, an evaluation was made by determining which topic was dominant for each record. To validate the results, the output gives the most appropriate topic for each text data. The topic ratios with their keywords can easily be seen. Through the model, a range of topics or recurring themes and the extent to which each document addresses the issues has been explored on the Kadi Registers.

## Findings and Discussion

The phrases used in the Kadi registers and the most frequently used words can be easily viewed in the graphic created by the LDA model. Indicators of time and place names and the name of the office where the court was held are among the most used words in the data set (Figure 15.2).

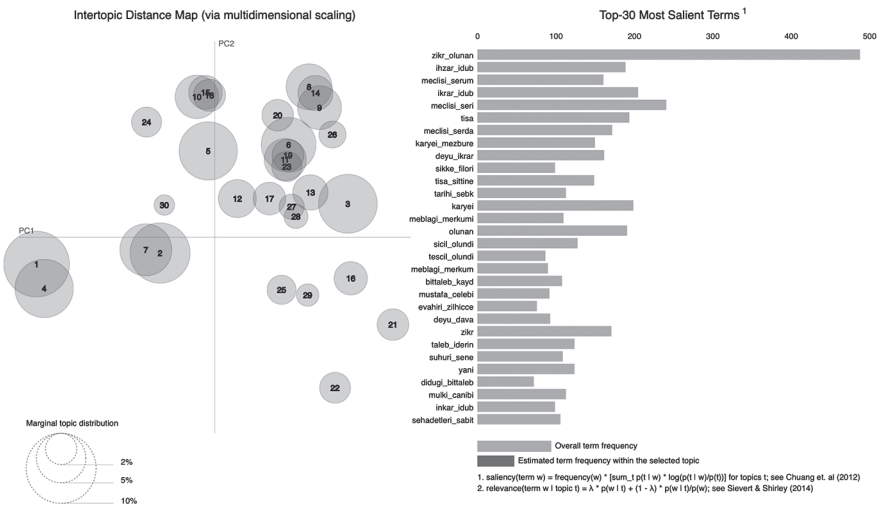


**Figure 15.2:** LDA algorithm on the Kadi Registers representing all clusters

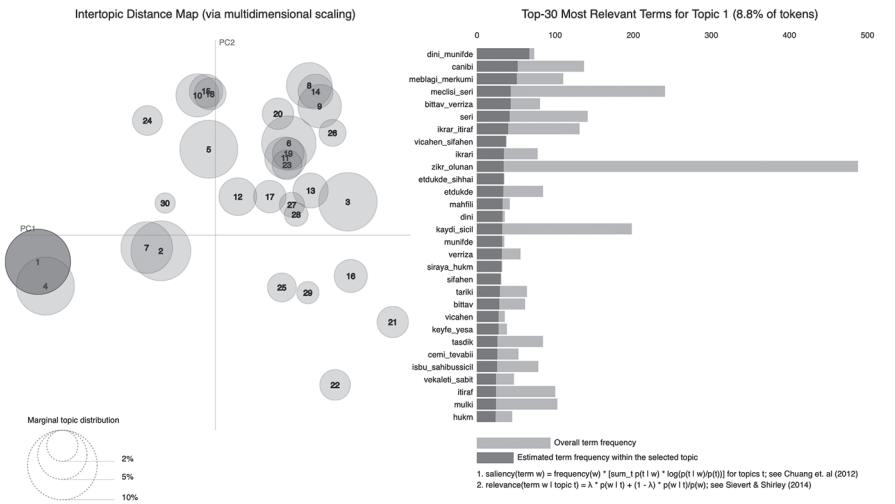


**Figure 15.3:** The first clusters of the LDA Model on the Kadi Registers

The first set of the LDA model shows that the most common records in the data set are *alım-satım (bey')*/commercial transaction cases which are represented by(the widest circle. The frequency of the words used in the first cluster in the whole document, red and blue, can be easily observed in Figure 15.3. Looking at the previous topic tagging, buying-selling cases constitute approximately 16% (101) of the total individual cases.

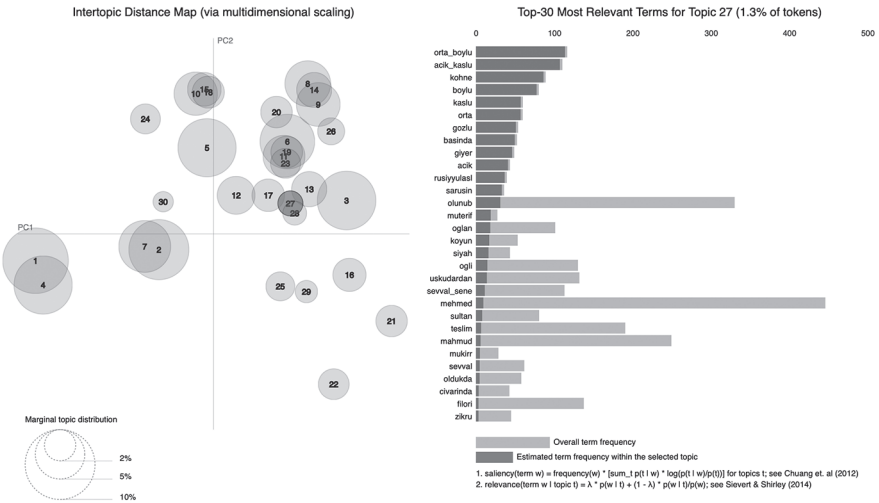


**Figure 15.4:** The 27th cluster of the LDA model on the Kadi Registers

In Figure 15.4, words used in the topic cluster, which are rarely mentioned in the whole data set, are seen. In a detailed reading, the topics of *iltizam*/tax farming, *kaçak köle*/fugitive slaves and *niza (kavga)*/ hostility) in the records of the Kadi registers were gathered in a cluster and indicate that the words used in the topics are related or that these topics can be included together under one record. In the LDA model figure, the cluster circles are smaller and the frequency of use of the words within this cluster in the entire data set has a distinctive character (Figure 15.4). The manual topic labelling made by the author on the Kadi register also confirms the outcomes. Accordingly, *iltizam*/tax farming is mentioned twenty-one times and the issue of *kaçak köle*/fugitive slaves forty times in the book (Akça 2005).

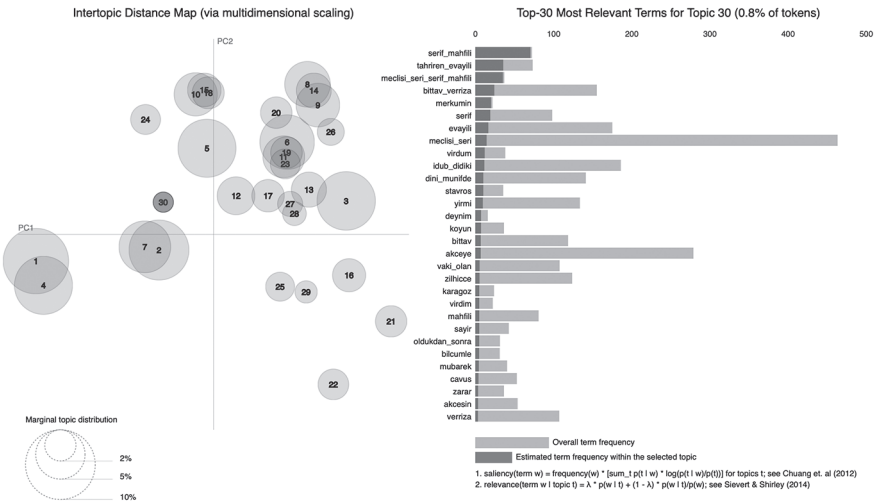


**Figure 15.5:** The 30th cluster of the LDA model on the Kadi Registers

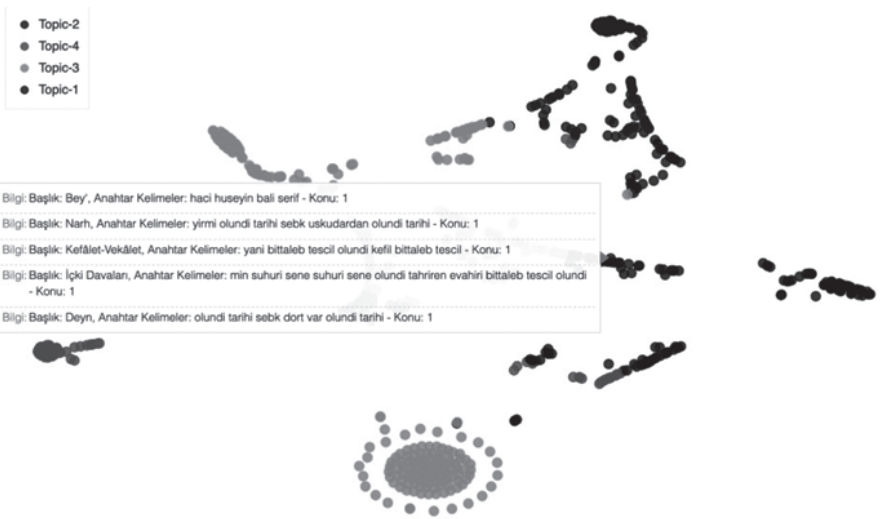


**Figure 15.6:** t-SNE algorithm on the dataset

The most distinctive issue in the data set is seen as the *kefil*/guarantor cases. Since the words used in *kefil*/guarantor or surety cases are generally names including foreign names and non- Muslim names, they have formed a separate cluster of words that are rarely used in the entire document universe (Figure 15.5). For a closer look at the LDA model and to examine each case, the t-distributed Stochastic

Neighbour Embedding (t-SNE) algorithm was applied to the data set. In Figure 15.6, the status of each case is visualised in two dimensions according to its topic. One of the important points is that the approach differentiates subjects which are unalike and unique in the data set and which subjects are similar to each other. Figure 15.6 identifies subjects which converge with each other and relate on the basis of easily used words. The words used in *alım-satım (bey'*)/commercial transaction, *borç (deyn)*/debt, *narh*/pricing, *kefil*/surety deputation and *alkol*/alcohol cases are seen as related words.

## Conclusion

The LDA topic modelling process used for analysing historical texts has emerged as an effective method that enables readers and researchers to comprehend content and identify subject matter for further examination. The results are presented in an efficient and contextualised way. AI and other means of automated analysis enhance the context and improve access to historical records. Computer methods provide significant improvements not only to the work of historians and social science researchers but also enhance social and cultural life and understandings. The results of this study can be used to guide future similar projects and also to enhance searching and information retrieval from the Kadi registers database and similar Ottoman document stores.

## Acknowledgement

I would like to thank Dr Müge Akbulut for her support in the study.

## References

Akbulut, Muge. 2022. "*Bilgi Erişimde İlgi Sıralamalarının Artırımlı Olarak Geliştirilmesi/* [Incremental Refinement of Relevance Rankings in Information Retrieval]." Unpublished doctoral dissertation, *Hacettepe Üniversitesi*, Ankara. https://bby.hacettepe.edu.tr/yayinlar/Muge_Akbulut_PhD_Tez.pdf

Akça, Sümeyye. 2005. "*Üsküdar Kadılığı 23 Nolu ve H. 968-970 Tarihli Sicilin Diplomatik Yönden Incelenmesi: Metin ve Inceleme*." Unpublished Master's thesis, *Marmara Üniversitesi*/Marmara University, İstanbul, Turkey. http://hdl.handle.net/11424/212437.

Andrés, José, José Ramón Prieto, Emilio Granell, Verónica Romero, Joan Andreu Sánchez, and Enrique Vidal. 2022, "Information Extraction from Handwritten Tables in Historical Documents." In

*Document Analysis Systems: Proceedings of the 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022*, edited by Seiichi Uchida, Elisa Barney, and Véronique Eglin, 184–198. Berlin: Springer Verlag. https://doi.org/10.1007/978-3-031-06555-2_13.

Ares Oliveira, Sofia, Benoit Seguin, and Frederic Kaplan. 2018. "DhSegment: A Generic Deep-learning Approach for Document Segmentation." In *Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR) Niagara Falls, NY, USA, 2018*, 7–12. doi: 10.1109/ICFHR-2018.2018.00011. Available at https://doi.org/10.1109/ICFHR-2018.2018.00011.

Arun, Rajkumar, et al. 2010. "On finding the natural number of topics with latent dirichlet allocation: Some observations." In *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14*. Springer Berlin Heidelberg.

Aydin, Bilgin, and Ekrem Tak. 2019. "Istanbul Sharia Court Registers." In *History of Istanbul* by the Türkiye Diyanet Foundation İslam Araştırmaları Merkezi/ Center for Islamic Studies (İSAM). Translated from the Turkish. Vol.2. https://istanbultarihi.ist/434-istanbul-sharia-court-registers?q=istanbul%20registers.

Bilgin Tasdemir, Esma F. 2023. "Printed Ottoman Text Recognition Using Synthetic Data and Data Augmentation." *International Journal on Document Analysis and Recognition* (IJDAR) 26: 273-287. https://doi.org/10.1007/s10032-023-00436-9. Available at https://assets.researchsquare.com/files/rs-2275909/v1_covered_f6102db4-105b-44e6-ad96-0713492be767.pdf?c=1697491252.

Cao, Juan, et al. 2009 "A density-based method for adaptive LDA model selection." *Neurocomputing* 72, no. 7–9: 1775–1781.

Craig, Lev, and David Petersson. 2023. "CNN vs. RNN: How Are They Different?" *TechTarget*, August 8, 2023. https://www.techtarget.com/searchenterpriseai/feature/CNN-vs-RNN-How-they-differ-and-where-they-overlap#:~:text=CNNs%20are%20commonly%20used%20to,such%20as%20text%20or%20videos.

Deveaud, Romain, Eric SanJuan, and Patrice Bellot. 2014. "Accurate and effective latent concept modeling for ad hoc information retrieval." *Document numérique* 17.1: 61-84.

Gao, Liangcai, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. 2019. "ICDAR 2019 Competition on Table Detection and Recognition (cTDaR)." In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 2019,* 1510–1515. doi: 10.1109/ICDAR.2019.00243.

Gilani, Azka, Shah Rukh Qasim, Imran Malik, and Faisal Shafait, 2017. "Table Detection Using Deep Learning" In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, 2017,* 771–776. https://tukl.seecs.nust.edu.pk/members/projects/conference/Table-Detection-Using-Deep-Learning.pdf.

Grant, Philip, Ratan Sebastian, Marc Allassonnière-Tang, and Sara Cosemans. 2021. "Topic Modelling on Archive Documents from the 1970s: Global Policies on Refugees." *Digital Scholarship in the Humanities* 36, no. 4: 886–904. https://doi.org/10.1093/llc/fqab018. Available at https://philip-grantlinguistics.net/docs/am_topic-modelling-dsh-2021.pdf.

Griffiths, Thomas L., and Mark Steyvers. 2004. "Finding scientific topics." *Proceedings of the National Academy of Sciences* 101, no. 1: 5228–5235. https://doi.org/10.1073/pnas.0307752101

Gültekin, Vedat. 2020. "Konu Otorite Dizini Nedir? Nasıl Oluşturulur?" *Türk Kütüphaneciliği/[Turkish Librarianship]* 34, no. 1: 46–64. https://dergipark.org.tr/tr/pub/tk/issue/53283/687062.

İstanbul Kadı Sicilleri. n.d. https://kadisicilleri.istanbul/.

Kirmizialtin, Suphan, and David Joseph Wrisley. 2022. "Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print Archive." *Digital Humanities Quarterly* 16, no. 2. https://www.digitalhumanities.org/dhq/vol/16/2/000577/000577.html.

Lang, Eva, Joan Puigcerver, Alejandro Héctor Toselli, and Enrique Vidal. 2018. “Probabilistic Indexing and Search for Information Extraction on Handwritten German Parish Records.” In *Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 2018*, 44–49. https://doi.org/10.1109/ICFHR-2018.2018.00017. Available at https://jpuigcerver.net/pubs/lang_icfhr2018.pdf.

*Luggat Osmanlica Türkçe Sözlük*/Luggat Ottoman Turkish Dictionary. 2024. Istanbul: EURODMC Dijital Marka Çözümleri Tic. Ltd. Şti. https://www.luggat.com/.

Motawa, Deya, Adnan Amin, and Robert Sabourin. 1997. “Segmentation of Arabic Cursive Script.” In *ICDAR ’97: Proceedings of the 4th International Conference on Document Analysis and Recognition, Ulm, Germany, August 1997*, 625–628. doi: 10.1109/ICDAR.1997.620580. Available at https://en.etsmtl.ca/ETS/media/ImagesETS/Labo/LIVIA/Publications/1997/sabourin97arabic.pdf.

Onat, Ayşe, Ferruh Yildiz, and Mesut Gündüz. 2008. “Ottoman Script Recognition Using Hidden Markov Model.” *International Journal of Computer and Information Engineering* 2, no. 2: 462–4. Available at https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f1d05b6257a661a788529f85579d39552f59391e.

Prasad, Animesh, Hervé Dejean, and Jean-Luc Meunier. 2019. “Versatile Layout Understanding Via Conjugate Graph.” In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Sydney, Australia, 20-25 September, 2019*, 287–294) doi: 10.1109/ICDAR.2019.00054. Available at https://animeshprasad.github.io/resources/c4.pdf.

Prieto, Jose Ramón, José Andrés, Emilio Granell, Joan Andreu Sánchez, and Enrique Vidal. 2023. “Information Extraction in Handwritten Historical Logbooks.” *Pattern Recognition Letters* 172: 128-136. https://doi.org/10.1016/j.patrec.2023.06.008.

Qasim, Shah Rukh, Hassan Mahmood, and Faisal Shafait. 2019. “ Rethinking Table Recognition Using Graph Neural Networks.” In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Sydney, Australia 2019*, 142–147. https://doi.ieeecomputersociety.org/10.1109/ICDAR.2019.00031. Available at https://doi.org/10.48550/arXiv.1905.13391.

Read Co-op. 2023. “Transkribus: Unlock Historical Documents with AI.” https://readcoop.eu/transkribus/.

*Sabancı Üniversitesi Dijital Beşeri Bilimler Laboratuvarı/* Sabanci University Digital Humanities Laboratory. n.d. “Akis: Ottoman Transcription Tool.” https://dhlab.sabanciuniv.edu/en/akis-osmanlica-transkripsiyon-araci.

Schöch, Christof. 2017. “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.” *Digital Humanities Quarterly* 11, no. 2. https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html. Also available at https://doi.org/10.48550/arXiv.2103.13019.

Siddiqui, Shoaib Ahmed, Imran Ali Fateh, Syed Tahseen Raza Rizvi, Andreas Dengel, Sheraz Ahmed. “DeepTabStR: Deep Learning Based Table Structure Recognition.” In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Sydney, Australia, 2019*, 1403–1409. doi: 10.1109/ICDAR.2019.00226. Available at https://www.dfki.de/fileadmin/user_upload/import/10649_DeepTabStR.pdf.

Wikipedia. 2024. “Wikilala.” Last updated February 24, 2024. https://en.wikipedia.org/wiki/Wikilala.

Yang, Tze-I, Andrew J. Torget, and Rada Mihalcea. 2011. “Topic Modeling on Historical Newspapers.” In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR. June 2011*, 96-104. Association for Computational Linguistics. https://aclanthology.org/W11-1513.pdf

Zhu, Scott, and Francois Chollet. 2023. ”Understanding Masking & Padding.” *TensorFlow Core Guide*, July 24, 2023. https://www.tensorflow.org/guide/keras/understanding_masking_and_padding.