OCR with Tesseract, Amazon Textract, and Google Document AI: A Benchmarking Experiment*

Thomas Hegghammer†

12 September 2021

Abstract

Optical Character Recognition (OCR) can open up understudied historical documents to computational analysis, but the accuracy of OCR software varies. This article reports a benchmarking experiment comparing the performance of Tesseract, Amazon Textract, and Google Document AI on images of English and Arabic text. English-language book scans (n=322) and Arabic-language article scans (n=100) were replicated 43 times with different types of artificial noise for a corpus of 18,568 documents, generating 51,304 process requests. Document AI delivered the best results, and the server-based processors (Textract and Document AI) performed substantially better than Tesseract, especially on noisy documents. Accuracy for English was considerably higher than for Arabic. Specifying the relative performance of three leading OCR products and the differential effects of commonly found noise types can help scholars identify better OCR solutions for their research needs. The test materials have been preserved in the openly available "Noisy OCR Dataset" (NOD) for reuse in future benchmarking studies.

Keywords: OCR, cloud computing, benchmarking

Word count: 8,313

*I am grateful to the three anonymous reviewers and to Neil Ketchley for valuable comments. I also thank participants in the University of Oslo Political Data Science seminar on 17 June 2021 for inputs and suggestions, as well as Eddie Antonio Santos for helping solve technical questions related to the ISRI/Ocreval tool. Supplementary information and replication materials are available at https://github.com/Hegghammer/noisy-ocr-benchmark.

†Norwegian Defence Research Establishment (FFI) - thomas.hegghammer@ffi.no

1 Introduction

Few technologies hold as much promise for the social sciences and humanities as optical character recognition (OCR). Automated text extraction from digital images can open up large quantities of understudied historical documents to computational analysis, potentially generating deep new insights on the human past.

But OCR is a technology still in the making, and available software provides varying levels of accuracy. The best results are usually obtained with a tailored solution involving corpus-specific pre-processing, model training, or postprocessing, but such procedures can be labour-intensive.[1] Pre-trained, general OCR processors have a much higher potential for wide adoption in the scholarly community, and hence their out-of-the box performance is of scientific interest.

For long, general OCR processors such as Tesseract (tesseract-ocr 2019; Patel, Patel, and Patel 2012) only delivered perfect results under what we may call laboratory conditions, i.e., on noise-free, single-column text in a clear printed font. This limited their utility for real-life historical documents, which often contain shading, blur, shine-through, stains, skewness, complex layouts, and other things that produce OCR error. Historically, general OCR processors have also struggled with non-Western languages (Kanungo, Marton, and Bulbul 1999), rendering them less useful for the many scholars

[1]For pre-processing see, e.g, Bieniecki, Grabowski, and Rozenberg (2007), Dengel et al. (1997), Holley (2009), Lat and Jawahar (2018), Volk, Furrer, and Sennrich (2011), and Wemhoener, Yalniz, and Manmatha (2013). For model training, see, e.g., Boiangiu et al. (2016), Reul et al. (2018), Springmann et al. (2014), and Wick, Reul, and Puppe (2018). For postprocessing, see, e.g., Kissos and Dershowitz (2016), Strohmaier et al. (2003), and Thompson, McNaught, and Ananiadou (2015).

working on documents in such languages.

In the past decade, advances in machine learning have led to substantial improvements in standalone OCR processor performance. Moreover, the past two years have seen the arrival of server-based processors such as Amazon Textract and Google Document AI, which offer document processing via an application processing interface (API) (Walker, Fujii, and Popat 2018). Media and blog coverage indicate that these processors deliver strong out-of-the-box performance[2], but those tests usually involve a small number of documents. Academic benchmarking studies exist (Tafti et al. 2016; Vijayarani and Sakila 2015) but the predate the server-based processors.

To find out, I conducted a benchmarking experiment comparing the performance of Tesseract, Textract, and Document AI on English and Arabic page scans. The objective was to generate statistically meaningful measurements of the accuracy of a selection of general OCR processors on document types commonly encountered in social scientific and humanities research.

The exercise yielded specifications for the relative performance of three leading OCR products as well as the differential effects of commonly found noise types. The findings can help scholars identify better OCR solutions for their research needs. The

[2]See, for example, Ted Han and Amanda Hickman, "Our Search for the Best OCR Tool, and What We Found," *OpenNews*, February 19, 2019 (https://source.opennews.org/articles/so-many-ocr-options/); Fabian Gringel, "Comparison of OCR tools: how to choose the best tool for your project," *Medium.com*, January 20, 2020 (https://medium.com/dida-machine-learning/comparison-of-ocr-tools-how-to-choose-the-best-tool-for-your-project-bd21fb9dce6b); Manoj Kukreja, "Compare Amazon Textract with Tesseract OCR — OCR & NLP Use Case," *TowardDataScience.com*, September 17, 2020 (https://towardsdatascience.com/compare-amazon-textract-with-tesseract-ocr-ocr-nlp-use-case-43ad7cd48748); Cem Dilmegani, "Best OCR by Text Extraction Accuracy in 2021," *AIMultiple.com*, June 6, 2021 (https://research.aimultiple.com/ocr-accuracy/).

test materials, which have been preserved in the openly available "Noisy OCR Dataset" (NOD), can be used in future research.

2 Design

The experiment involved taking two document collections of 322 English-language and 100 Arabic-language page scans, replicating them 43 times with different types of artificially generated noise, processing the full corpus of ~18,500 documents in each OCR engine, and measuring the accuracy against ground truth using the Information Science Research Institute (ISRI) tool.

2.1 Processors

I chose Tesseract, Textract, and Document AI on the basis of their wide use, reputation for accuracy, and availability for programmatic use. Budget constraints prevented the inclusion of additional reputable processors such as Adobe PDF Services and ABBYY Cloud OCR, but these can be tested in the future using the same procedure and test materials.[3]

A full description of these processors is beyond the scope of this article, but Table 1 summarizes their main user-related features.[4] All the processors are primarily

[3]As of September 2021, Adobe PDF Services charges a flat rate of $50 per 1,000 pages (https://www.adobe.io/apis/documentcloud/dcsdk/pdf-pricing.html, accessed 3 September 2021). ABBYY Cloud costs between $28 and $60 per 1,000 pages depending one's monthly plan and the total number of documents (see https://www.abbyy.com/cloud-ocr-sdk/licensing-and-pricing/, accessed 3 September 2021). By contrast, processing in Amazon Textract and Google Document AI costs $1.50 per 1,000 pages.

[4]For documentation, see the product websites: https://github.com/tesseract-ocr/tesseract,

Table 1: Features of Tesseract, Textract, and Document AI

Name	Maintainer	Installation	Architecture	Languages	Cost
Tesseract	Tesseract OCR Project	Local	LSTM	116	Free
Textract	Amazon Web Services	Server-based	Undisclosed	6	$1.50 per 1000 pages
Document AI	Google Cloud Services	Server-based	Undisclosed	60+	$1.50 per 1000 pages

designed for programmatic use and can be accessed in multiple programming languages, including R and Python. The main difference is that Tesseract is open source and installed locally, whereas Textract and Document are paid services accessed remotely via a REST API.

2.2 Data

For test data, I sought materials that would be reasonably representative of those commonly studied in the social sciences and humanities. This is to say historical documents containing extended text, as opposed to forms, receipts, and other business documents, which commercial OCR engines are primarily designed for, and which tend to get the most attention in media and blog reviews.

Since many scholars work on documents in languages other than English, I also wanted to include test materials in a non-Western language. Historically, these have been less well served by OCR engines, partly because their sometimes more ornate scripts are more difficult to process than Latin script, and partly because market incentives have led the software industry to prioritize the development of English-language OCR. I chose Arabic for three reasons: its size as a world language, its

https://aws.amazon.com/textract/, and https://cloud.google.com/document-ai.

alphabetic structure (which allows accuracy measurement with the ISRI tool), and the complexity of its script. Arabic is known as one of the hardest alphabetic languages for computers to process (Mariner 2017; Jain, Mathew, and Jawahar 2017), so including it alongside English will likely provide something close to the outer performance bounds of OCR engines on alphabetic scripts. I excluded logographic scripts such as Hanzi (Chinese) and Kanji (Japanese) partly due to the difficulty of generating comparable accuracy measures and partly due to my lack of familiarity with such languages.

The English test corpus consisted of the "Old Books Dataset" (Barcha 2017), a collection of 322 colour page scans from ten books printed between 1853 and 1920 (see figures 1a and 1b and Table 2). The dataset comes as 300 DPI and 500 DPI TIFF image files accompanied by ground truth (drawn from the Project Gutenberg website) in TXT files. I used the 300 DPI files in the experiment.

The Arabic test materials were drawn from the "Yarmouk Arabic OCR Dataset" (Doush, AlKhateeb, and Gharibeh 2018), a collection of 4,587 Wikipedia articles printed out to paper and colour scanned to PDF (see figures 1c and 1d). The dataset contains ground truth in HTML and TXT files. Due to the homogeneity of the collection, a randomly selected subset of 100 pages was deemed sufficient for the experiment.

The Yarmouk dataset is suboptimal because it does not come from historical printed documents, but it is one of very few Arabic language datasets of some size with accompanying ground truth data. The English and Arabic test materials are

thus not directly analogous, and in principle the latter poses a lighter OCR challenge than the former. Another limitation of the experiment is that the test materials only includes single-column text due to the complexities involved in measuring layout parsing accuracy.

2.3 Noise application

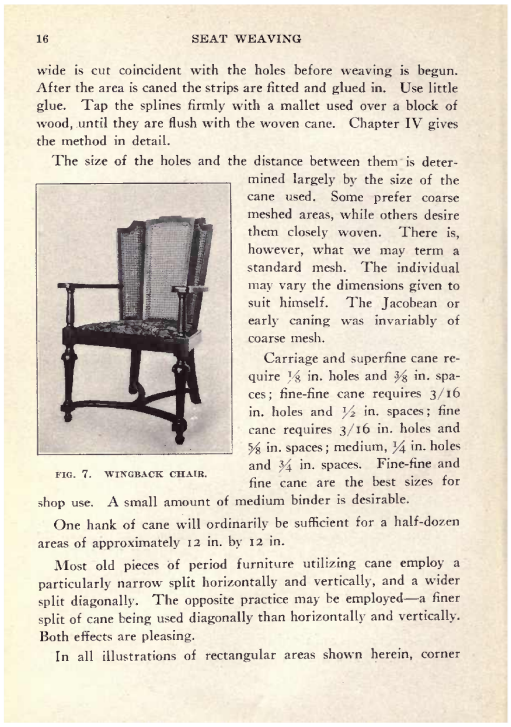
Social scientists and historians often deal with digitized historical documents that contain visual noise (Krishnan and Babu 2012; Ye and Doermann 2013). In practice, virtually any document that existed first on paper and were later digitized — which is to say almost all documents produced before around 1990 and many thereafter — is going to contain some kind of noise. Sometimes it is the original copy that is degraded; at other times the document passed through a poor photocopier, an old microfilm, or a blurry lens before reaching us. The type and degree of noise will vary across collections and individual documents, but most scholars who use archival material will encounter this problem at least occasionally.

A key objective of the experiment was therefore to gauge the effect of different types of visual noise on OCR performance. To achieve this, I programmatically applied different types of artifical noise to the test materials, so as to allow isolation of noise effects at the measurement stage. Specifically, the two dataset were duplicated 43 times, each with a different type of noise filter. The R code used for noise generation is included in the Appendix.[5]

[5]There are other ways of generating synthetic noise, notably the powerful tool DocCreator (Journet

Figure 1: Sample test documents in their original state

(a) Old Books j020



16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

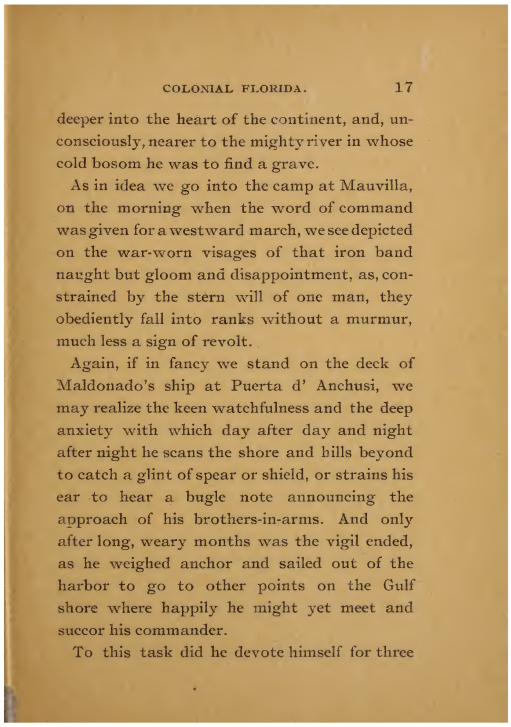
Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ½ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(b) Old Books g023



COLONIAL FLORIDA. 17

deeper into the heart of the continent, and, unconsciously, nearer to the mighty river in whose cold bosom he was to find a grave.

As in idea we go into the camp at Mauvilla, on the morning when the word of command was given for a westward march, we see depicted on the war-worn visages of that iron band naught but gloom and disappointment, as, constrained by the stern will of one man, they obediently fall into ranks without a murmur, much less a sign of revolt.

Again, if in fancy we stand on the deck of Maldonado's ship at Puerta d' Anchusi, we may realize the keen watchfulness and the deep anxiety with which day after day and night after night he scans the shore and hills beyond to catch a glint of spear or shield, or strains his ear to hear a bugle note announcing the approach of his brothers-in-arms. And only after long, weary months was the vigil ended, as he weighed anchor and sailed out of the harbor to go to other points on the Gulf shore where happily he might yet meet and succor his commander.

To this task did he devote himself for three

(c) Yarmouk 25223-1

إحيائية - ويكيبيديا، الموسوعة الحرة

الإحيائية (بالأنجليزية: Animism، انيميزم بمعنى "الروح، حية" من اللاتينية)[1][2] هي الاعتقاد بوجود الأرواح وأن أي نظام حي أو كائن أو حتى المواد الجامدة أحياناً تمتلك نوعاً من الروح مثل الحجارة والنباتات. وكذلك في الظواهر الطبيعية مثل الرعد. وحتى داخل المعالم جغرافية كالجبال والأنهار.[3] الإحيائية منتشرة في الديانات القومية، ولكن يمكن العثور عليها في الشنتو، والهندوسية، ووحدة الوجود، والسيخية، والوثنية الجديدة.

على مر تاريخ أوروبا، كان العديد من الفلاسفة قد فكروا باحتمال وجود الروح في الحيوان، الأسنان، والنبات، ولكن التصور المقبول الحالي قد وضعه إدوارد بيرنت تايلور في القرن التاسع عشر.[4]

على الرغم من وجه الشبه بين الطوطمية والإحيائية، ولكنها كما يقول تيم انجواد تركز على كائنات الروح الفردية التي تساعد في إدامة الحياة، في حين الطوطمية تقول أن هذاك مصدر يوفر الأساس للحياة مثل الأرض، أو أسلاف. بعض الجماعات الدينية القومية مثل السكان الأصليين لاسترالية يميلون إلى الطوطمية، ولكن الإويت يميلون إلى الإحيائية في نظرهم.[5]

مصطلح الإحيائية أتى من كلمة *animismus* التي وضعت من قبل العالم الألماني إدوارد بيرنت تايلور في حوالي 1720م للإشارة إلى أن "يتم انتاج الحياة الحيوانية من قبل روح معنوية". الانيميزم" دخل اللغة الإنجليزية في عام 1819م.[6] وقد اتخذ المصطلح وقاموا بإعادة تعريفه من قبل الموارد تايلور الأنثروبولوجي في كتابه"الثقافة البدائية" عام 1871م، حرفه ب"النظرية الحركة المفاهيمية الطبيعية". بفضل تعريف تايلور له، ينظر الاحيائيون إلى العالم الطبيعي على أنه على قيد الحياة. الوصف الذاتي لنفسه كان أنه "عالم عقلاني معزز"، كان يعتقد أن هذا الرأي كان "طفولي" ونموذج "التخلف المعرفي".[7] وبالتالي إنه كان شائعة في الشعوب "البدائية" مثل أولئك الذين يعيشون في مجتمعات جامعي الصيد.

مفهوم تايلور للإحيائية استُخدم من قبل العديد من علماء الأنثروبولوجيا، مثل إميل دوركهايم وكلود ليفي ستروس. وتيم أنجواد. ولكن بعض علماء الأنثروبولوجيا، مثل نوريت بيرت،دنيفن قد انتقدوا مفهوم تايلور للإحيائية، واعتبروه أنه قديم.[8]

الإحيائية بشكل عام هو على سبيل المثال اعتبار الأجسام على أنها تحتوي على روح ومعاملتهم كما أن الأجسام أرواح على نحو شبه عالمي. جان بياجيه طبق المصطلح في علم نفس الطفل في الإشارة إلى تفاهم التضمني بالعالم في عقل الطفل الذي يقترض أن كل الأحداث تنتج من نية أو وعي. بياجيه يفسر هذا بعدم القدرة المعرفية للتمييز بين العالم الخارجي من نفسية المرء. علم النفس التنموي قد قرر ان التمييز الأشكال ذوات الأرواح مثقال الجماد أنها فكرة مجردة المكتسبة عن طريق التعلم.

توانم الظواهر التي أدت الوحشيين إلى الاعتقاد بوجود الإحيائية أعطيت من قبل تيلور، هربرت سبنسر، أندرو لانج. وغيرهم؛ قد اتسمت جدل إحيائي بين المضيئية والقوائم الأولوية. ومن هذه الظواهر، الأحلام والهلوسات النشوة.

في العادة أي مجموعة من المعتقدات يروج الطبيعة المختلفة قد تكون موصوفة بأنها"احيائية"، ولكن هذه العبارة سبب الكثير من الاختلافات بالرأي بين علماء الثقافة الأنثروبولوجيا. وصف الديانات التقليدية الأفريقية ب"الإحيائية" يعتبر كمصطلح يستخدم لاختصار التفاصل الأعلى والأكثر تعقيدا بين الشيوخ، والأجداد ورمح الطبيعة.

الشنتو، التي تعد الديانة التقليدية في اليابان، تعد إحيائية للغاية. في الشنتو، روح الطبيعة، أو الكامي، موجود في كل مكان، من الاعظم (مثل آلهة الشمس)، الذي يمكن اعتبار ها تعدد، إلى الأدنى، التي قد تصف بشكل من أشكال الإحيائية.

هناك بعض الجماعات الهندوسية التي يمكن أن تعتبر إحيائية. في ولاية كارناتاكا الساحلية لديهم عادة مختلفة للصلاة للأرواح لإلههم.

الكثير من الديانات التقليدية الأصلية الأمير يكية في الأساس إحيائية. انظر، على سبيل المثال، لصلاة اكوتا سوز ميثاكويه أويباسين. صلاة دعى الشكر لدى الإيروكواس تشكر كل الامي. نبات، وحيوان.

العديد لكن ليس كل ديانات الوثنية الجديدة يمارسون شكل من أشكال الإحيائية.

(d) Yarmouk 4155-1

نقيل التاج (اب) - ويكيبيديا، الموسوعة الحرة

نقيل التاج محلة تابعة لقرية علوان، التابعة لعزلة بني مدسم بمديرية اب إحدى مديريات محافظة إب في الجمهورية اليمنية[1]، بلغ تعداد سكانها 45 حسب تعداد اليمن لعام 2004.

الأحكري - الأسهوم - الأكاسى - الأكمة - البرعات - الجالة - الجاح، الجاح، إب - الجاشة - الجبانة - الجبب - الجبجبة - الجحلة - الجديد - الجرابي - الجرجرة - الجرابين - الجماشي - الجميمة - الجوز - الجير - الحجر - الحجلة - الحجفة، إب - الحرجة - الحصن - الحصن، إب - الحنامي - الحوالين - الحيمة - الخرابة - الخربة - الخربه، إب - الخرشية - الخيراف - الدثلة - الدخلة - الدفيقة - الدنوه - الذيلية - الذنيبة - الربعاء - الربيض - الرجمة - الساكن - السر - السرة - السعيدية - السلق - السلف - السماري - الشهلة - الصواري - الضبر - الشحاف - الشرنمة - الشعب الداخل - الشعب الكبير - الشماحي - الشهاص - الصالحية - الصباحي - الصباص - الشبة - الصبر - الصنبرة - الضحيكة - الضرائم - الضلضف - الضوالع - الطاهر - الطرانمة - الظهرة - العارضة - العارضه، إب - العمهلي - العوف - العوالق - العرقلي - العريقي، إب - العلوف - العكلة - العلاجية - العوادي - الغراب - الفلق - الفلعاتي - القبلي - القبلي، إب - القرانبي - القرات - القلعة - القيف - الكحلاني - الكريف - الدلاح - المادح - المجدية - المجمعة - المجيل - المحاشة - المحايط - المحلاق - المحجر - المحدادة - المحرفة - المحطابة - المحل - المدغة - المرابض - المرزوم - المرزوقة - المسوح - المسوح، إب - المشاعبة - المشاعبه، إب - المشراق، إب - المشعب - المصالعه - المعاين - المعبر - المعبرين - المعطن - المعين - المغسل - المقارع - المناس - المناخ - المناخ، إب - المنجية - المنظر - المنقل - النافش - النجد - النوبة - الهيجة - الوالي - الهارى - اليهاري - يعسنة - بيت الجيل - بيت السريحي - بيت الغيلان - بيت سعيد ناصر - بيت مراد - تبان - جبات - جبل برط - جبل مشرقي - جعران - جوابة العليا - جوابة السفلى - حصباية - حصن بن شايع - حطلة - حير - خيالة - خرانب نصر - دار الغيل - دار المرية - دار عماري - دارمرشد - ذعلة - ذى حجنب - ذى حمادة - ذى شعيرة - ذى شيبة - ذى صبيلة - ذى عجزب - ذى كريب - ذى محباش - ذي حبيش - ذي سريرة - ذي سود - رجاح - سالله العين - سكران - سواد - سواد اب - سوق الثلاث - شاكان - شحنان - شعب الجمل - شعب الضيف - شعبه الرمادي - صبيل الكبرة - صرحدد - طلبة - ظهره السادة - عدن الذيب - عرائج - عرش القيف - غرض الجدوب - عشاي - عقد علوان - عليان - فراد - فروع - قاحدان السفلى - قاحدان العليا - قريضة العليا - قريضه السفلى - قصالة - كسلية - مانر - محجر السفي - محضرة - مشورة - مصريات - معري - معرية - منزل الركة - منزل الشعراني - منزل حرف - منزل موسى - منقذة - موران - نقيل الدوري - نقيل السنف - نوابه السفلى - نوابه العليا - هدام - هير - الهقاير - بيوت العدن - نعيمة

I began by creating a binary version of each image, so that there were two versions — colour and greyscale — with no added noise (see figure 2a and 2b). I then wrote functions to generate six ideal types of image noise: "blur," "weak ink," "salt and pepper," "watermark," "scribbles," and "ink stains" (see figures 2c-h). While not an exhaustive list of possible noise types, they represent several of the most common ones found in historical document scans.[6] I applied each of the six filters to both the colour version and the binary version of the images, thus creating 12 additional versions of each image. Lastly I applied all available combinations of two noise filters to the colour and binary images, for an additional 30 versions.

This generated a total of 44 image versions divided into three categories of noise intensity: 2 versions with no added noise, 12 versions with one layer of noise, and 30 versions with two layers of noise. This amounted to an English test corpus of 14,168 documents and an Arabic test corpus of 4,400 documents. The dataset is preserved as the "Noisy OCR Dataset" (Hegghammer 2021).

et al. 2017). I chose not to use DocCreator primarily because it is graphical user interface-based, and I found I could generate realistic noise more efficiently with R code.

[6]It would be possible to extend the list of noise types further, to include 10-20 different types, but this would increase the size of the corpus (and thus the processing costs) considerably, probably without affecting the broad result patterns. Since the main aim here is not to map all noise types but to compare processors, I decided on a manageable subset of noise types.

Figure 2: Sample test document ("Old Books j020") with noise applied

(a) Original

16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

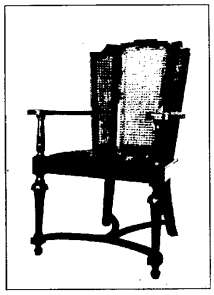
Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ¼ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(b) Binary



16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ¼ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(c) Blur

16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ¼ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(d) Weak ink

16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ¼ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(e) Salt and pepper

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ½ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(f) Watermark

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ½ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(g) Scribbles

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ½ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

(h) Ink stains

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

FIG. 7. WINGBACK CHAIR.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3/16 in. holes and ½ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

2.4 Processing

The experiment aimed at measuring out-of-the-box performance, so documents were submitted without further preprocessing using the OCR engines' default settings.[7] While this is an uncommon use of Tesseract, it treats the engines equally and helps highlight the degree to which Tesseract is dependent on image preprocessing.

The English corpus was submitted to all three OCR engines in a total of 42,504 document processing requests. The Arabic corpus was only submitted to Tesseract and Document AI — since Textract does not support Arabic — for a total of 8,800 processing requests.

The Tesseract processing was done in R with the package `tesseract` (v4.1.1). For Textract, it was carried out via the R package `paws` (v0.1.11), which provides a wrapper for the Amazon Web Services API. For Document AI, I used the R package `daiR` (v0.8.0) to access the Document AI API v1 endpoint. The processing was done in April and May of 2021 and took an estimated net total of 150-200 hours to complete. The Document AI and Textract APIs processed documents at a rate of approximately 10-15 seconds per page. Tesseract took 17 seconds per page for Arabic and 2 seconds per page for English on a Linux Desktop with a 12-core, 4.3 Ghz CPU and 64GB RAM.

[7] The only exception was the setting of the relevant language libraries in Tesseract.

2.5 Measurement

Accuracy was measured with the ISRI tool (Rice and Nartker 1996) in Eddie Antonio Santos's (2019) updated version — known as Ocreval — which has UTF-8 support. ISRI is a simple but robust tool that has been used for OCR assessment since its creation in the mid-1990s. Alternatives exist (Alghamdi, Alkhazi, and Teahan 2016; Carrasco 2014; Yalniz and Manmatha 2011), but ISRI was deemed sufficient for this exercise.

ISRI compares two versions of a text — in this case OCR output to ground truth — and returns a range of measures for divergence, notably a document's overall character accuracy and word accuracy expressed in percent. Character accuracy is the proportion of characters in a hypothesis text that match the reference text. Any misread, misplaced, absent, or excess character is considered an error and subtracted from the numerator. This represents the so-called Levenshtein distance (Levenshtein and others 1966), i.e., the minimum number of edit operations needed to correct the hypothesis text. Word accuracy is the proportion of non-stopwords in a hypothesis text that match those of the reference text.[8]

Character and word accuracy are usually highly correlated, but the former punishes error harder, since each wrong character detracts from the accuracy rate.[9] In word

[8]ISRI only has an English-language stopword list (of 110 words), so in the measurements for Arabic, stopwords are included in the assessment. All else equal, this should produce slightly higher accuracy rates for Arabic, since oft-recurring words are easier for OCR engines to recognize.

[9]ISRI's character accuracy rates can actually be negative as a result of excess text. OCR engines sometimes introduce garbled text when they see images or blank areas with noise, resulting in output texts that are much longer than ground truth. Since excess characters are treated as errors and subtracted from the numerator, they can result in negative accuracy rates. In the corpus studied

accuracy, by contrast, a misspelled word counts as one error regardless of the number of wrong characters that contribute to the error. Moreover, in ISRI's implementation of word accuracy, case errors and excess words are ignored.[10]

Figure 3 provides some examples of what character and word error rates may correspond to in an actual text. I will return later to the question of how error matters for analysis.

Which of the two measures is better depends on the type of document and the purpose of the analysis. For shorter texts where details matter — such as forms and business documents — character accuracy is considered the more relevant measure. For longer texts to be used for searches or text mining, word accuracy is commonly used as the principal metric. In the following, I therefore report word accuracy rates, transformed to word error rates by subtracting them from 100. Character accuracy rates are available in the Appendix.

3 Results

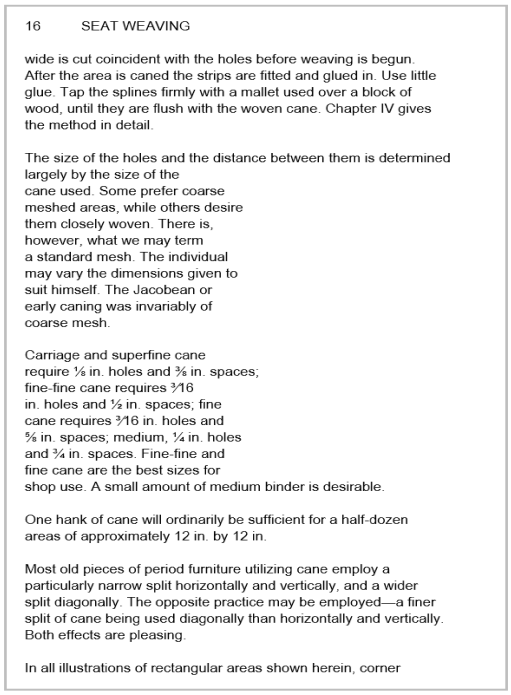
The main results are shown in Figure 4 and reveal clear patterns. Document AI had consistently lower error rates, with Textract coming in a close second, and

here, this phenomenon affected 4.6 percent of the character accuracy measurements, and it occurred almost exclusively in texts processed by Tesseract.

[10]This also means that ISRI's word accuracy tool does not yield negative rates. As Eddie Antonio Santos explains, "The wordacc algorithm creates parallel arrays of words and checks only for words present in the ground truth. It finds 'paths' from the generated file that correspond to ground truth. For this reason, it only detects as many words as there are in ground truth"; private email correspondence, 1 September 2021. However, the word accuracy tool returns NA when the hypothesis text has no recognizable words. This occurred in 9.4 percent of the measurements in this experiment, again almost exclusively in Tesseract output. These NAs are treated as zeroes in figures 4 to 6.

Figure 3: Examples of word error effects

(a) Ground truth



16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

Carriage and superfine cane require ⅛ in. holes and ⅜ in. spaces; fine-fine cane requires 3⁄16 in. holes and ⅛ in. spaces; fine cane requires 3⁄16 in. holes and ⅝ in. spaces; medium, ¼ in. holes and ¾ in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

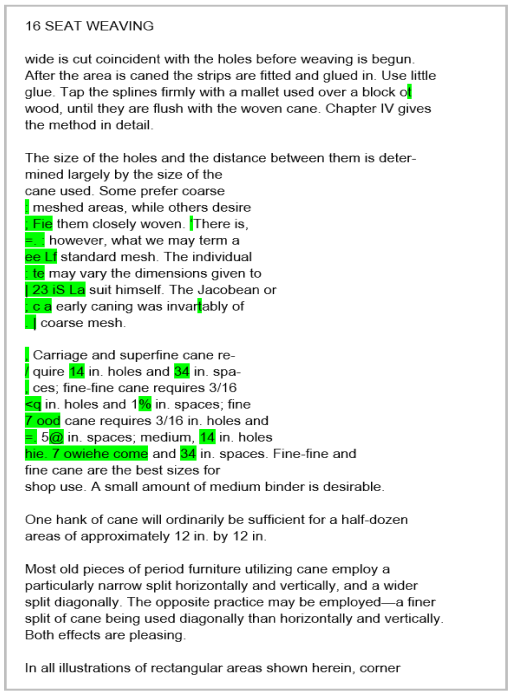
One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

Old Books j020

(b) ~5% word error (~8% character error)



16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. Some prefer coarse meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual may vary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

Carriage and superfine cane require 14 in. holes and 34 in. spaces; fine-fine cane requires 3/16 in. holes and 1⅛ in. spaces; fine cane requires 3/16 in. holes and ⅝ in. spaces; medium, 14 in. holes and 34 in. spaces. Fine-fine and fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

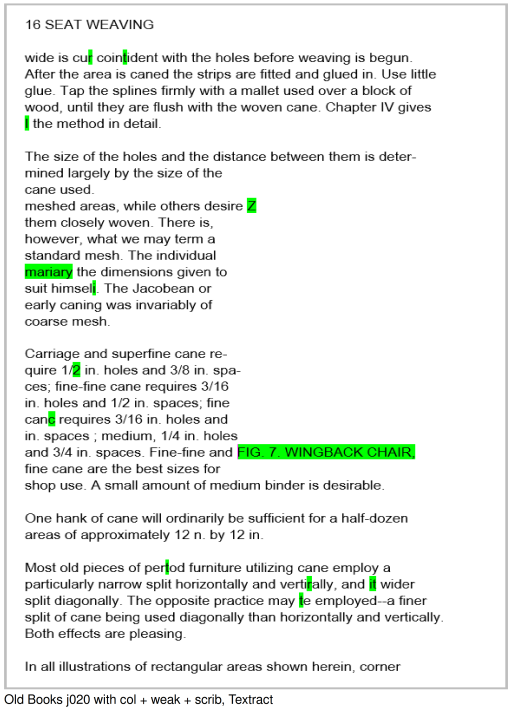
One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 in. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and a wider split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

Old Books j020 with col + wm, Tesseract

(c) ~10% word error (~8% character error)



16 SEAT WEAVING

wide is cut coincident with the holes before weaving is begun. After the area is caned the strips are fitted and glued in. Use little glue. Tap the splines firmly with a mallet used over a block of wood, until they are flush with the woven cane. Chapter IV gives the method in detail.

The size of the holes and the distance between them is determined largely by the size of the cane used. meshed areas, while others desire them closely woven. There is, however, what we may term a standard mesh. The individual mariary the dimensions given to suit himself. The Jacobean or early caning was invariably of coarse mesh.

Carriage and superfine cane require 1/2 in. holes and 3/8 in. spaces; fine-fine cane requires 3/16 in. holes and 1/2 in. spaces; fine canc requires 3/16 in. holes and in. spaces ; medium, 1/4 in. holes and 3/4 in. spaces. Fine-fine and FIG. 7. WINGBACK CHAIR. fine cane are the best sizes for shop use. A small amount of medium binder is desirable.

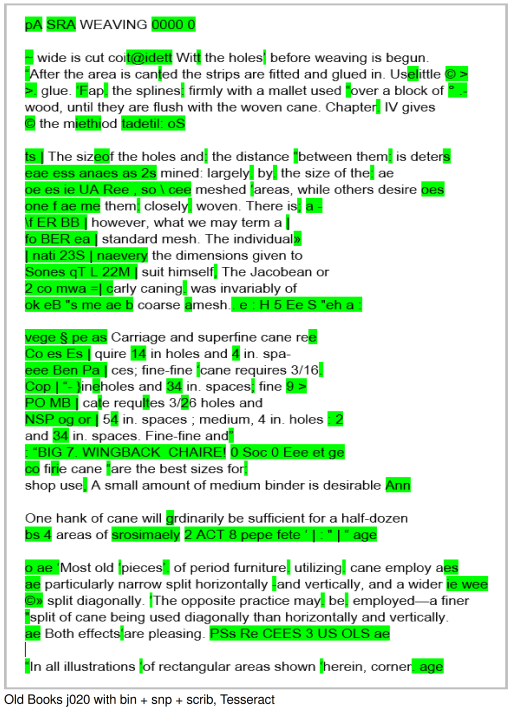
One hank of cane will ordinarily be sufficient for a half-dozen areas of approximately 12 n. by 12 in.

Most old pieces of period furniture utilizing cane employ a particularly narrow split horizontally and vertically, and ll wider split diagonally. The opposite practice may le employed--a finer split of cane being used diagonally than horizontally and vertically. Both effects are pleasing.

In all illustrations of rectangular areas shown herein, corner

Old Books j020 with col + weak + scrib, Textract

(d) ~20% word error (~37% character error)



pA SRA WEAVING 0000 0

= wide is cut coit@idett Wit the holes before weaving is begun. After the area is canled the strips are fitted and glued in. Uselittle © > >. glue. Tap the splines firmly with a mallet used over a block of " - wood, until they are flush with the woven cane. Chapter IV gives © the methiod tadetil. oS

ts. The sizedf the holes and the distance between them is deters eae ess anaes as 2s mined: largely by the size of the ae oe es ie UA Ree ., so \ cee meshed areas, while others desire oes one f.ae me them closely woven. There is a - If ER BB. however, what we may term a fo BER ea standard mesh. The individual nati 23S naevery the dimensions given to Sones qT L 22M suit himself The Jacobean or 2 co mwa =| darly caning was invariably of ok eB "s me ae b coarse amesh. e : H 5 Ee S "eh a

vege § pe as Carriage and superfine cane ree Co es Es quire 14 in holes and 4 in. spaeee Ben Pa ces; fine-fine cane requires 3/16 Cop inholes and 34 in. spaces fine 9 > PO MB cale requires 3/26 holes and NSP og or 54 in. spaces ; medium, 4 in. holes 2 and 34 in. spaces. Fine-fine and "BIG 7. WINGBACK CHAIRE 0 Soc 0 Eee et ge cq fine cane are the best sizes for shop use A small amount of medium binder is desirable Ann

One hank of cane will ordinarily be sufficient for a half-dozen bs 4 areas of srosimately 2 ACT 8 pepe fete ' | : " | " age

o ae Most old pieces of period furniture utilizing cane employ aes ae particularly narrow split horizontally and vertically, and a wider le wee ow split diagonally. The opposite practice may be employed—a finer split of cane being used diagonally than horizontally and vertically. ae Both effects are pleasing. PSs Re CEES 3 US OLS ae

In all illustrations of rectangular areas shown herein, corner age

Old Books j020 with bin + snp + scrib, Tesseract

Errors hand annotated in green. Omitted text unmarked.

Tesseract last. More noise yielded higher error rates in all engines, but Tesseract was significantly more sensitive to noise than the two others. Overall, there was a significant performance gap between the server-based processors (Document AI and Textract) on one side and the local installation (Tesseract) on the other. Only on noise-free documents in English could Tesseract compete.

We also see a marked performance difference across languages. Both Document AI and Tesseract delivered substantially lower accuracy for Arabic than they did for English. This was despite the Arabic corpus consisting of Internet articles in a single, very common font, while the English corpus contained old book scans in several different fonts. An analogous Arabic corpus would likely have produced an even larger performance gap. This said, Document AI represents a significant improvement on Tesseract as far as out-of-the-box Arabic OCR is concerned.

Disaggregating the data by noise type shows a more detailed picture (see Figures 5 and 6). Beyond the patterns already described, we see, for example, that both Textract and Tesseract performed somewhat better on greyscale versions of the test images than on the colour version. We also note that all engines struggled with blur, while Tesseract was much more sensitive to salt & pepper noise than the two other engines. Incidentally, it is not surprising that the ink stain filter yielded lower accuracy throughout since it completely concealed part of the text. The reason we see a bimodal distribution in the bin + blur" filters on the English corpus is that they yielded many zero values, probably as a result of the image crossing a threshold of illegibility. The

same did not happen in the Arabic corpus, probably because the source images there had crisper characters at the outset.

Figure 4: Word error rates by engine and noise level for English and Arabic documents

Mean error rates in coloured boxes. p < .05 in Kolmogorov−Smirnov tests for all distribution pairs. X axes cropped.

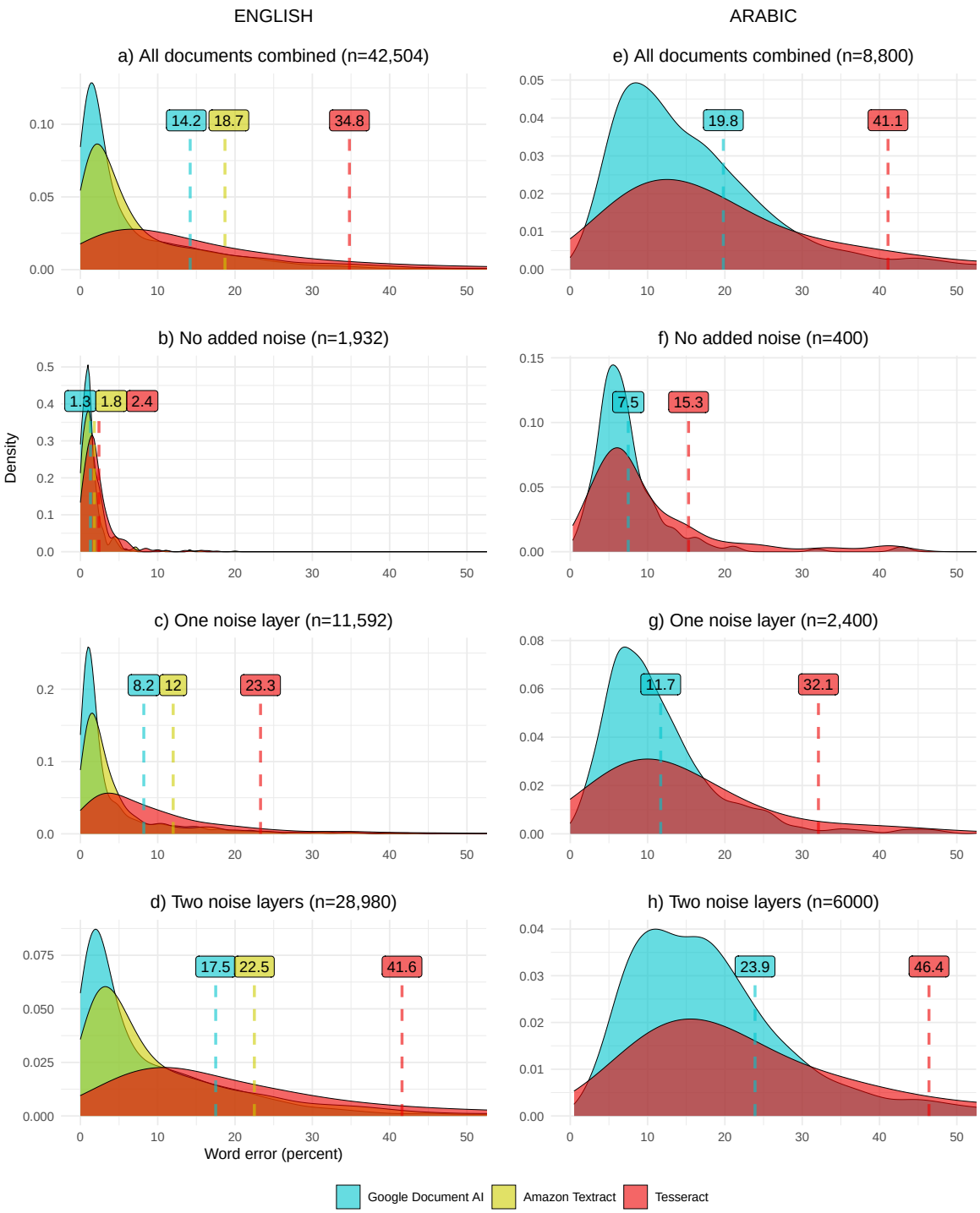


Figure 5: Word error rates by engine and noise type for English-language documents

Data: Single−column text in historical book scans with noise added artificially (n=42,504; 322 per engine and noise type).
Noise codes: 'col'=colour, 'bin'=binary, 'blur'=blur, 'weak'=weak ink, 'snp'=salt&pepper, 'wm'=watermark, 'scrib'=scribbles, 'ink'=ink stains.

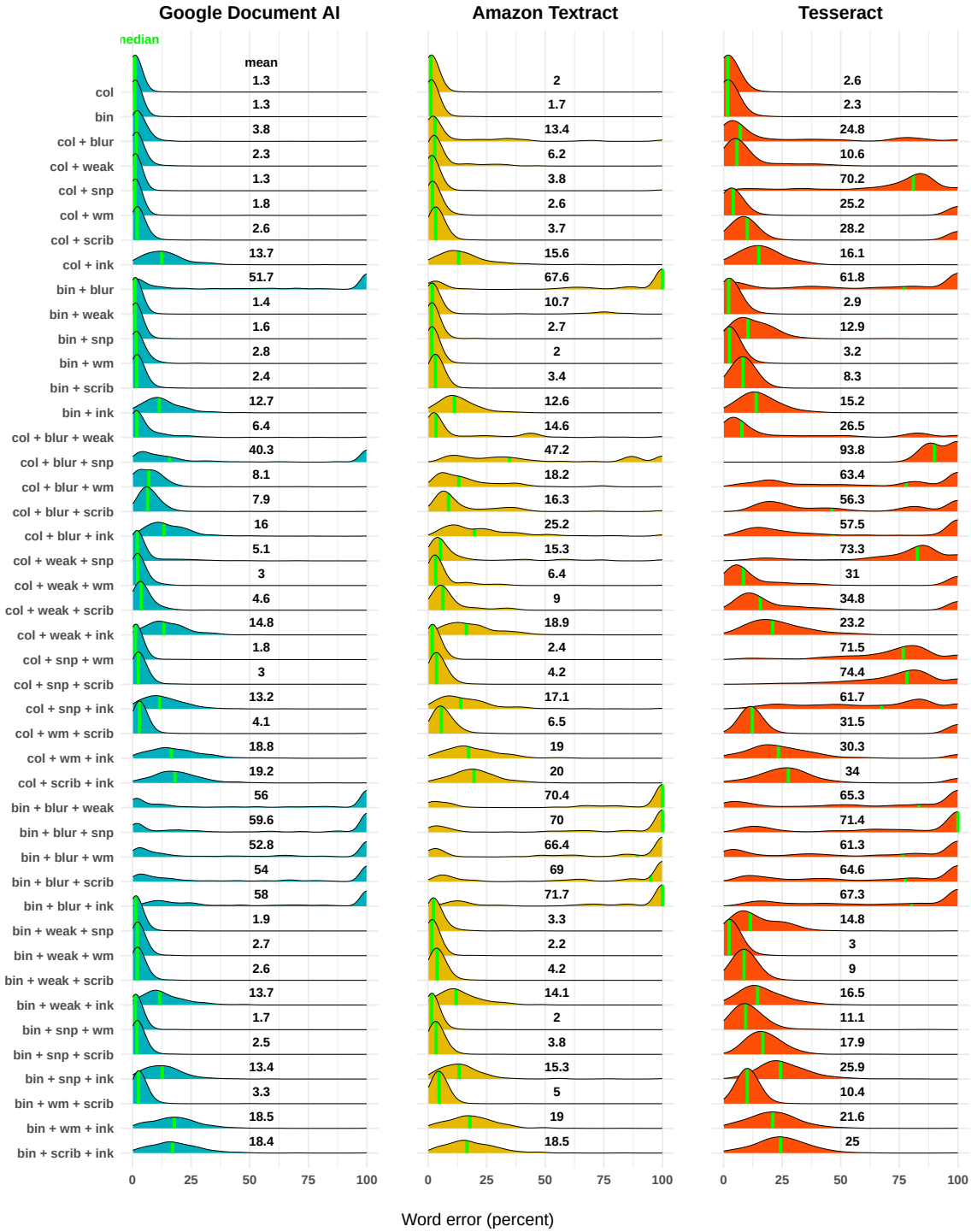
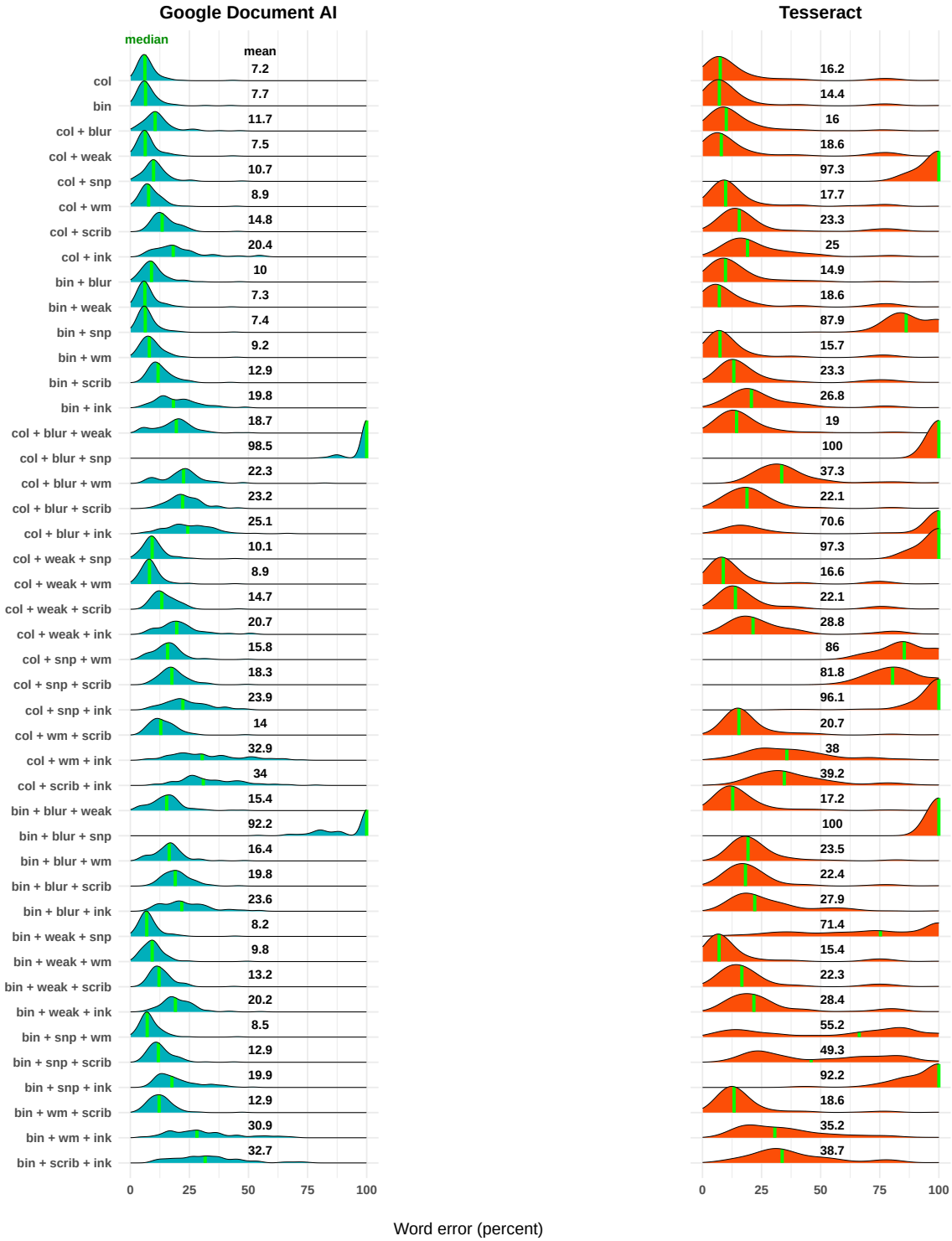


Figure 6: Word error rates by engine and noise type for Arabic-language documents

Data: Single−column text in image scans of Arabic Wikipedia pages with noise added artificially (n = 8800; 100 per engine and noise type).
Noise codes: 'col'=colour, 'bin'=binary, 'blur'=blur, 'weak'=weak ink, 'snp'=salt&pepper, 'wm'=watermark, 'scrib'=scribbles, 'ink'=ink stains.



4 Implications

When is it worth paying for better OCR accuracy? The answer depends on a range of situational factors, such as the state of the corpus, the utility function of the researcher, and the intended use case.

Much hinges on the corpus itself. As we have seen, accuracy gains increase with noise and are higher for certain types of noise. Moreover, if the corpus contains many different types of noise, a better general processor will save the researcher relatively more preprocessing time. Unfortunately we lack good tools for (ground truth-free) noise diagnostics, but there are ways to obtain some information about the noise state of the corpus (Gupta et al. 2015; Lins, Banergee, and Thielo 2010; Reffle and Ringlstetter 2013). Finally, the size of the dataset matters, since processing costs scale with the number of documents while accuracy gains do not.

The calculus also depends on the economic situation of the researcher. Aside from absolute size of one's budget, a key consideration is labour cost, since cloud-based processing is in some sense a substitute for Tesseract processing with additional labour input. The latter option will thus make more sense for a student than for a professor and more sense for the faster programmer.

Last but not least is the intended use of the OCRed text. If the aim is to recreate a perfect plaintext copy of the original document for, say, a browseable digital archive, then every percentage point matters. But if the purpose is to build a topic model or conduct a sentiment analysis, it is not obvious that a cleaner text will always

yield better end results. The downstream effects of OCR error is a complex topic that cannot be explored in full here, but we can get some pointers by looking at the available literature and doing some tests of our own.

Existing research suggests that the effects of OCR error vary by analytical toolset. Broadly speaking, topic models have proved relatively robust to OCR inaccuracy (Colavizza 2021; Grant et al. 2021; Mutuvi et al. 2018; Su et al. 2015), with Van Strien et al (2020) suggesting a baseline for acceptable OCR accuracy as low as 80 percent. Classification models have been somewhat more error-sensitive, although the results here have been mixed (Colavizza 2021; Murata et al. 2006; Stein, Argamon, and Frieder 2006; van Strien. et al. 2020). The biggest problems seem to arise in natural language processing (NLP) tasks where details matter, such as part-of-speech tagging and named entity recognition (Hamdi et al. 2019; Lopresti 2009; Miller et al. 2000; van Strien. et al. 2020).

To illustrate some of these dynamics and add to the empirical knowledge of OCR error effects, we can run some simple tests on the English-language materials from our benchmarking exercise. The Old Books dataset is small, but similar in kind to the types of text collections studied by historians and social scientists, and hence a reasonably representative test corpus. In the following, I look at OCR error in four analytical settings: sentiment analysis, classification, topic modelling, and named entity recognition. I exploit the fact that the benchmarking exercise yielded 132 different variants (3 engines and 44 noise types) of the Old Book corpus, each with

a somewhat different amount of OCR error.[11] By running the same analyses on all text variants, we should get a sense of how OCR error can affect substantive findings. This said, the exercise as a whole is a back-of-the-envelope test insofar as it covers only a small subset of available text mining methods and does not implement any of them as fully as one would in a real-life setting.

4.1 Sentiment analysis

Faced with a corpus like Old Books (see Table 2), a researcher might want to explore text sentiment, for example to examine differences between authors or over time. Using the R package `quanteda`s LSD 2015 and ANEW dictionaries, I generated document-level sentiment polarity and valence scores for all variants of the corpus after standard preprocessing. To assess the effect of OCR error, I calculated the absolute difference between these scores and those of the ground truth version of the corpus. Figures 7a-d indicate that these differences increase only slightly with OCR error, but also that, for sentiment polarity, the variance is such that just a few percent OCR error can produce sentiment scores that diverge from ground truth scores by up to two whole points at the document level.

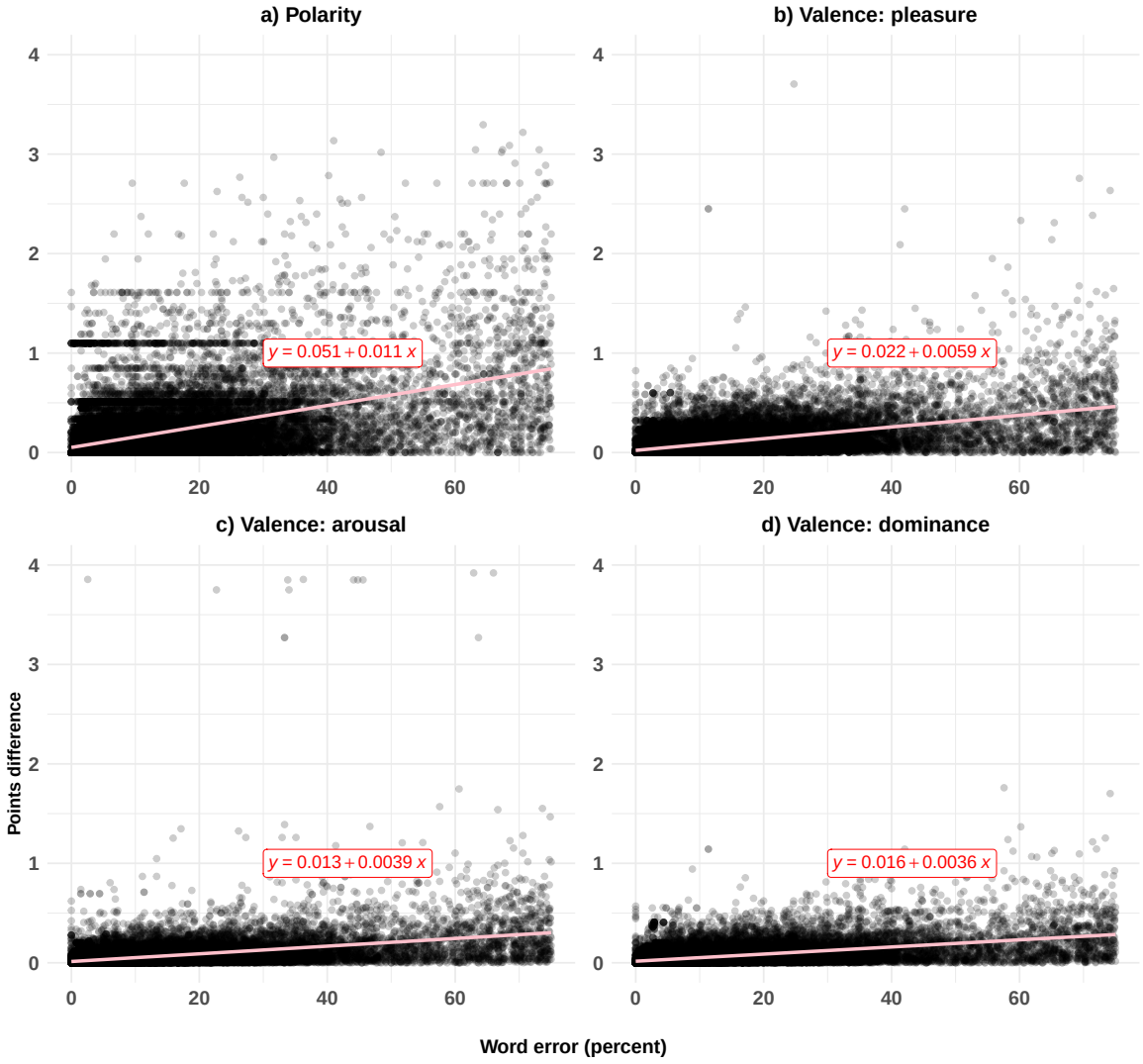
[11] In all of the below, "OCR error" refers to word error rates computed with the ISRI tool.

Table 2: Composition of Old Books corpus

Title	Author	Year	Pages	Words
Engraving of Lions, Tigers, Panthers, Leopards, Dogs, &C.	Thomas Landseer	1853	8 (28)	3983
The Corset and the Crinoline	William Barry Lord	1868	30 (254)	9633
Horton Genealogy	George Firman Horton	1876	34 (316)	11744
Historical Sketches of Colonial Florida	Richard Lewis Campbell	1892	30 (284)	4801
Half-Hours with the Highwaymen	Charles George Harper	1908	34 (422)	7695
Betrayed Armenia	Diana Agabeg Apcar	1910	39 (77)	15001
The Lusitania's Last Voyage	Charles Emelius Lauriat, Jr.	1915	23 (162)	3438
The Child of the Moat	Ian B. Stoughton Holborn	1916	30 (408)	7844
Seat Weaving	L. Day Perry	1917	57 (96)	12437
The Boy Apprenticed to an Enchanter	Padraic Colum	1920	37 (168)	7420

Figure 7: OCR error and sentiment analysis accuracy

Differences in document−level (n=42504) sentiment polarity and valence between OCR processed versions of 'Old Books' and ground truth. Scores calculated with Quanteda's 'LSD 2015' and 'ANEW' dictionaries. Y axis is absolute difference in polarity/valence points from ground truth score. X axis cropped.



4.2 Text classification

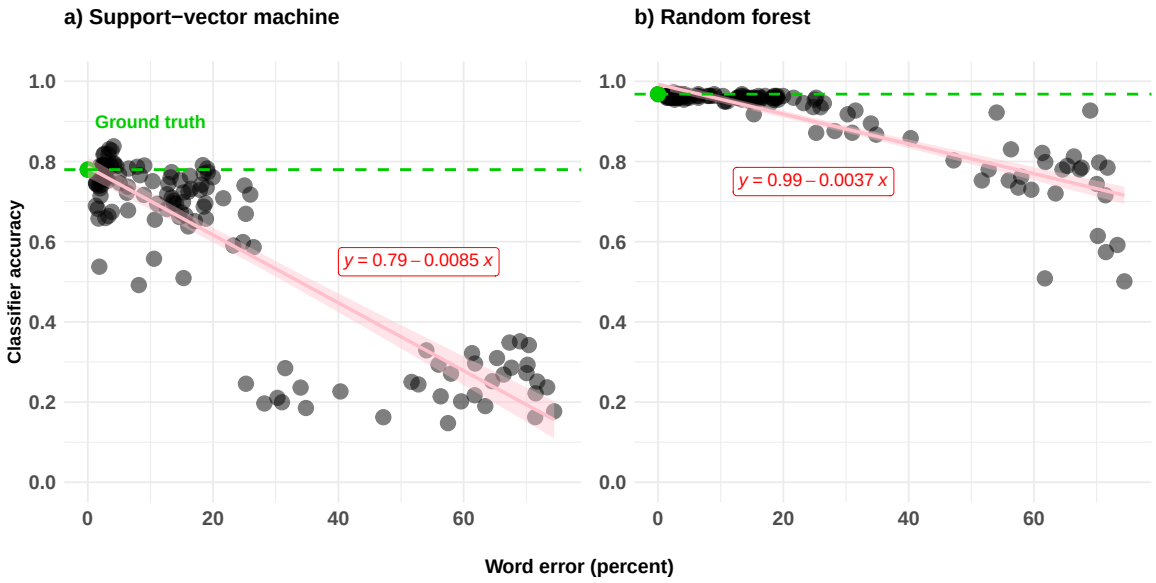
Another common analytical task is text classification. Imagine that we knew which works were represented in the Old Books corpus, but not which work each document belonged to. We could then handcode a subset and train an algorithm to classify the rest. Since we happen to have pre-coded metadata we can easily simulate this exercise. I trained two multiclass classifiers — Random Forest and Support-Vector Machine — to retrieve the book from which a document was drawn. To avoid imbalance, I removed the smallest subset ("Engraving of Lions, Tigers, Panthers, Leopards, Dogs, &C.") and was left with 9 classes and 314 documents. For each variant of the corpus I preprocessed the texts, split them 70/30 for training and testing, and fit the models using the `tidymodels` R package. Figures 8a and 8b show the results. We see that OCR error has only a small negative effect on classifier accuracy up to a threshold of around 20 percent OCR error, after which accuracy plummets.

4.3 Topic modelling

Assessing the effect of OCR error on topic models is more complicated, since they involve more judgment calls and do not yield an obvious indicator of accuracy. I used the `stm` R package to fit structural topic models to all the versions of the corpus. As a first step, I ran the `stm::searchK()` function for a k value range from 6 to 20, on the suspicion that different variants of the text might yield different diagnostics and hence inspire different choices for the number of topics in the model. Figure 9a shows that

Figure 8: OCR error and multiclass classifier accuracy

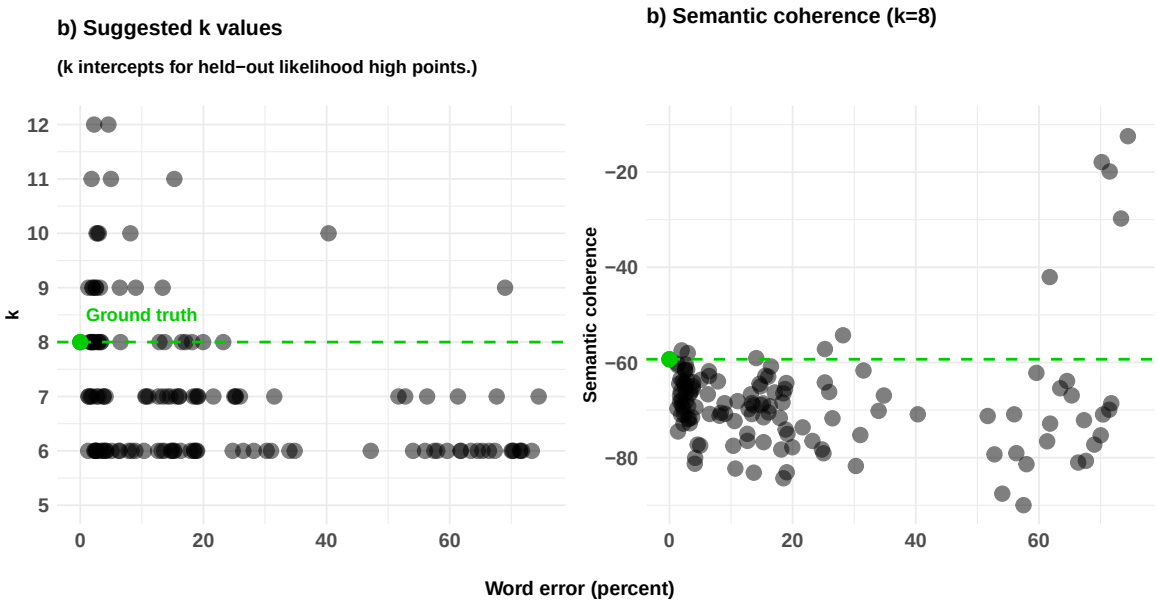
Classifiers trained on Old Books dataset (9 classes). Each point represents a noise/engine version (n=132) of the corpus. X axis cropped.



the k intercept for the high point of the held-out likelihood curve varies from 6 to 12 depending on the version of the corpus. Held-out likelihood is not the only criterion for selecting k, but it is an important one, so these results suggests that even a small amount of OCR error can lead researchers to choose a different topic number than they would have done on a cleaner text, with concomitant effects on the substantive analysis. Moreover, if we hold k still at 8 — the value suggested by diagnostics of the ground truth version of the corpus — we see in Figure 9b that the semantic coherence of the model decreases slightly with more noise.

Figure 9: OCR error and topic model fits

Structural topic models of Old Books dataset, built with R package stm. Each point represents a noise/engine version (n=132) of the corpus. X axes cropped.

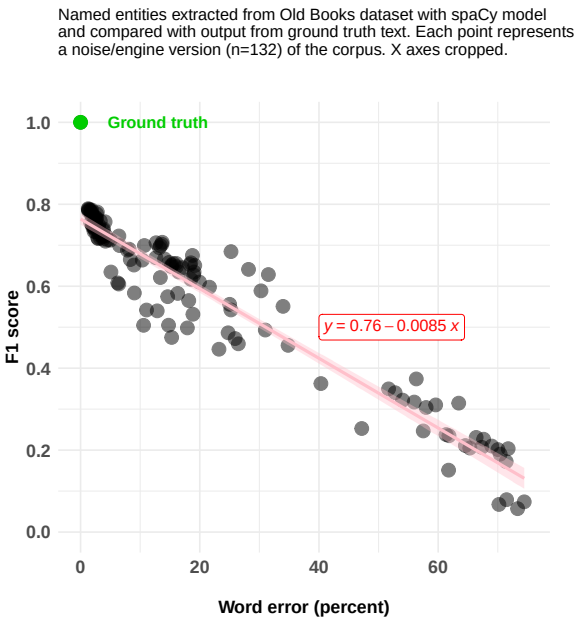


4.4 Named entity recognition

Our corpus is full of names and dates, so a researcher might also want to explore it with named entity recognition (NER) models. I used a pretrained `spaCy` model (`en_core_web_sm`) to extract entities from all non-preprocessed versions of the corpus and compared the output to that of the ground truth text. In the absence of ground truth NER label data, I treated `spaCy`'s prediction for the ground truth text as the reference point and calculated the F1 score (the harmonic average of precision and recall) as a metric for accuracy. For simplicity, the evaluation included only predicted entity names, not entity labels. Figure 10 shows that OCR error affected NER accuracy severely. In a real-life setting these effects would be partly mitigated by pre- and

postprocessing, but it seems reasonable to suggest that NER is one of the areas where the value added from high-precision OCR is the highest.

Figure 10: OCR error and named entity recognition accuracy



Broadly speaking, these tests indicate that OCR error mattered the most in NER, the least in topic modelling and sentiment analysis, while in classification there was a tipping point at around 20 percent OCR error. At the same time, all the tests showed some accuracy deterioration even at very low OCR error rates.

5 Conclusion

This article described a systematic test of three general OCR processors on a large new dataset of English and Arabic documents. It suggests that the server-based engines Document AI and Textract deliver markedly higher out-of-the-box accuracy

than the standalone Tesseract library, especially on noisy documents. It also indicates that certain types of "integrated" noise, such as blur and salt and pepper, generate more error than "superimposed" noise such as watermarks, scribbles, and even ink stains. Furthermore, it suggests that the "OCR language gap" still persists, although Document AI seems to have partially closed it, at least for Arabic.

The key takeaway for the social sciences and humanities is that high-accuracy OCR is now more accessible than ever before. Researchers who might be deterred by the prospect of extensive document preprocessing or corpus-specific model training now have at their disposal user-friendly tools that deliver strong results out of the box. This will likely lead to more scholars adopting OCR technology and to more historical documents becoming digitized.

The findings can also help scholars tailor OCR solutions to their needs. For many users and use cases, server-based OCR processing will be an efficient option. However, there are are downsides to consider, such as processing fees and data privacy concerns, which means that in some cases, other solutions — such as self-trained Tesseract models or even plain Tesseract — might be preferable.[12] Having baseline data on relative processor performance and differential effects of noise types can help navigate such tradeoffs and optimise one's workflow.

[12]Amazon openly says it "may store and use document and image inputs [...] to improve and develop the quality of Amazon Textract and other Amazon machine-learning/artificial-intelligence technologies" (see https://aws.amazon.com/textract/faqs/, accessed 3 September 2021). Google says it "does not use any of your content [...] for any purpose except to provide you with the Document AI API service" (see https://cloud.google.com/document-ai/docs/data-usage, accessed 3 September 2021), but it is unclear what lies in the word "provide" and whether it includes the training of the processor.

The study has several limitations, notably that it tested only three processors on two languages with a non-exhaustive list of noise types. This means we cannot say which processor is the very best on the market or provide a comprehensive guide to OCR performance on all languages and noise types. However, the test design used here can easily be applied to other processors, languages, and noise types for a more complete picture. Another limitation is that the experiment only used single-column test materials, which does not capture layout parsing capabilities. Most OCR engines, including Document AI and Textract, still struggle with multi-column text, and even state-of-the-art tools such as Layout Parser (Shen et al. 2021) require corpus-specific training for accurate results. Future studies will need to determine which processors deliver the best out-of-the-box layout parsing. In any case, we appear to be in the middle of a small revolution in OCR technology with potentially large benefits for the social sciences and humanities.

6 Conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alghamdi, Mansoor A, Ibrahim S Alkhazi, and William J Teahan. 2016. "Arabic OCR Evaluation Tool." In *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, 1–6. IEEE.

Barcha, Pedro. 2017. "Old Books Dataset." *GitHub Repository*. GitHub. https://github.com/PedroBarcha/old-books-dataset.

Bieniecki, Wojciech, Szymon Grabowski, and Wojciech Rozenberg. 2007. "Image Preprocessing for Improving Ocr Accuracy." In *2007 International Conference on Perspective Technologies and Methods in MEMS Design*, 75–80. IEEE.

Boiangiu, Costin-Anton, Radu Ioanitescu, Razvan-Costin Dragomir, and others. 2016. "Voting-Based OCR System." *The Proceedings of Journal ISOM* 10: 470–86.

Carrasco, Rafael C. 2014. "An Open-Source OCR Evaluation Tool." In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 179–84.

Colavizza, Giovanni. 2021. "Is your OCR good enough? Probably so. Results from an assessment of the impact of OCR quality on downstream tasks." *KB Lab Blog*.

https://lab.kb.nl/about-us/blog/your-ocr-good-enough-probably-so-results-assessment-impact-ocr-quality-downstream.

Dengel, Andreas, Rainer Hoch, Frank Hönes, Thorsten Jäger, Michael Malburg, and Achim Weigel. 1997. "Techniques for Improving OCR Results." In *Handbook of Character Recognition and Document Image Analysis*, 227–58. World Scientific.

Doush, Iyad Abu, Faisal AlKhateeb, and Anwaar Hamdi Gharibeh. 2018. "Yarmouk Arabic OCR Dataset." In *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, 150–54. IEEE.

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*, March. https://doi.org/10.1093/llc/fqab018.

Gupta, Anshul, Ricardo Gutierrez-Osuna, Matthew Christy, Boris Capitanu, Loretta Auvil, Liz Grumbach, Richard Furuta, and Laura Mandell. 2015. "Automatic Assessment of OCR Quality in Historical Documents." In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1.

Hamdi, Ahmed, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2019. "An Analysis of the Performance of Named Entity Recognition over OCRed Documents." In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 333–34. IEEE. https://ieeexplore.ieee.org/document/8791217.

Hegghammer, Thomas. 2021. "Noisy OCR Dataset." Repository details TBC.

Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3/4).

Jain, Mohit, Minesh Mathew, and C. V. Jawahar. 2017. "Unconstrained Scene Text and Video Text Recognition for Arabic Script." http://arxiv.org/abs/1711.02396.

Journet, Nicholas, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. 2017. "Doccreator: A New Software for Creating Synthetic Ground-Truthed Document Images." *Journal of Imaging* 3 (4): 62.

Kanungo, Tapas, Gregory A Marton, and Osama Bulbul. 1999. "Performance Evaluation of Two Arabic OCR Products." In *27th AIPR Workshop: Advances in Computer-Assisted Recognition*, 3584:76–83. International Society for Optics; Photonics.

Kissos, Ido, and Nachum Dershowitz. 2016. "OCR Error Correction Using Character Correction and Feature-Based Word Classification." In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 198–203. IEEE.

Krishnan, R, and DR Ramesh Babu. 2012. "A Language Independent Characterization of Document Image Noise in Historical Scripts." *International Journal of Computer Applications* 50 (9): 11–18.

Lat, Ankit, and CV Jawahar. 2018. "Enhancing Ocr Accuracy with Super Resolution." In *2018 24th International Conference on Pattern Recognition (ICPR)*, 3162–67. IEEE.

Levenshtein, Vladimir I, and others. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." In *Soviet Physics Doklady*, 10:707–10. 8. Soviet Union.

Lins, Rafael Dueire, Serene Banergee, and Marcelo Thielo. 2010. "Automatically Detecting and Classifying Noises in Document Images." In *Proceedings of the 2010 ACM Symposium on Applied Computing*, 33–39.

Lopresti, Daniel. 2009. "Optical Character Recognition Errors and Their Effects on Natural Language Processing." *International Journal on Document Analysis and Recognition (IJDAR)* 12 (3): 141–51. http://www.cse.lehigh.edu/~lopresti/tmp/AND08journal.pdf.

Mariner, Matthew C. 2017. "Optical Character Recognition (OCR)." In *Encyclopedia of Computer Science and Technology*, 622–29. CRC Press.

Miller, David, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 2000. "Named Entity Extraction from Noisy Input: Speech and OCR." In *Sixth Applied Natural Language Processing Conference*, 316–24. https://aclanthology.org/A00-1044.pdf.

Murata, Mayo, Lazaro S. P. Busagala, Wataru Ohyama, Tetsushi Wakabayashi, and Fumitaka Kimura. 2006. "The Impact of OCR Accuracy and Feature Transformation on Automatic Text Classification." In *Document Analysis Systems VII*, edited by Horst Bunke and A. Lawrence Spitz, 506–17. Berlin, Heidelberg: Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/11669487_45.

Mutuvi, Stephen, Antoine Doucet, Moses Odeo, and Adam Jatowt. 2018. "Evaluating the Impact of OCR Errors on Topic Modeling." In *International Conference on Asian Digital Libraries*, 3–14. Springer.

Patel, Chirag, Atul Patel, and Dharmendra Patel. 2012. "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study." *International Journal of Computer Applications* 55 (10): 50–56.

Reffle, Ulrich, and Christoph Ringlstetter. 2013. "Unsupervised Profiling of OCRed Historical Documents." *Pattern Recognition* 46 (5): 1346–57.

Reul, Christian, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. "Improving OCR Accuracy on Early Printed Books by Utilizing Cross Fold Training and Voting." In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 423–28. IEEE.

Rice, Stephen V, and Thomas A Nartker. 1996. "The ISRI Analytic Tools for OCR Evaluation." *UNLV/Information Science Research Institute, TR-96* 2.

Santos, Eddie Antonio. 2019. "OCR Evaluation Tools for the 21st Century." In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, 23–27. Honolulu: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-6004.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. "LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis." *arXiv Preprint arXiv:2103.15348*.

Springmann, Uwe, Dietmar Najock, Hermann Morgenroth, Helmut Schmid, Annette Gotscharek, and Florian Fink. 2014. "OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress." In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 71–75.

Stein, Sterling Stuart, Shlomo Argamon, and Ophir Frieder. 2006. "The Effect of OCR Errors on Stylistic Text Classification." In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 701–2. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.6791&rep=rep1&type=pdf.

Strohmaier, Christian M, Christoph Ringlstetter, Klaus U Schulz, and Stoyan Mihov. 2003. "Lexical Postcorrection of OCR-Results: The Web as a Dynamic Secondary Dictionary?" In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 3:1133–33. Citeseer.

Su, Jing, Oisin Boydell, Derek Greene, and Gerard Lynch. 2015. "Topic Stability over Noisy Sources." *arXiv Preprint arXiv:1508.01067*. https://noisy-text.github.io/2016/pdf/WNUT09.pdf.

Tafti, Ahmad P, Ahmadreza Baghaie, Mehdi Assefi, Hamid R Arabnia, Zeyun Yu, and Peggy Peissig. 2016. "OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym." In *International Symposium on Visual Computing*, 735–46. Springer.

tesseract-ocr. 2019. "Tesseract OCR 4.1.1." *GitHub Repository*. GitHub. https://github.com/tesseract-ocr/tesseract.

Thompson, Paul, John McNaught, and Sophia Ananiadou. 2015. "Customised OCR Correction for Historical Medical Text." In *2015 Digital Heritage*, 1:35–42. IEEE.

van Strien., Daniel, Kaspar Beelen., Mariona Ardanuy., Kasra Hosseini., Barbara McGillivray., and Giovanni Colavizza. 2020. "Assessing the Impact of OCR Quality on Downstream NLP Tasks." INSTICC; SciTePress. https://doi.org/10.5220/0009169004840496.

Vijayarani, S, and A Sakila. 2015. "Performance Comparison of OCR Tools." *International Journal of UbiComp (IJU)* 6 (3): 19–30.

Volk, Martin, Lenz Furrer, and Rico Sennrich. 2011. "Strategies for Reducing and Correcting OCR Errors." In *Language Technology for Cultural Heritage*, 3–22. Springer.

Walker, Jake, Yasuhisa Fujii, and Ashok C Popat. 2018. "A Web-Based Ocr Service for Documents." In *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, Austria*. Vol. 1.

Wemhoener, David, Ismet Zeki Yalniz, and R Manmatha. 2013. "Creating an Improved Version Using Noisy OCR from Multiple Editions." In *2013 12th International Conference on Document Analysis and Recognition*, 160–64. IEEE.

Wick, Christoph, Christian Reul, and Frank Puppe. 2018. "Comparison of OCR Accuracy on Early Printed Books Using the Open Source Engines Calamari and

OCRopus." *J. Lang. Technol. Comput. Linguistics* 33 (1): 79–96.

Yalniz, Ismet Zeki, and R. Manmatha. 2011. "A Fast Alignment Scheme for Automatic OCR Evaluation of Books." In *2011 International Conference on Document Analysis and Recognition*, 754–58. https://doi.org/10.1109/ICDAR.2011.157.

Ye, Peng, and David Doermann. 2013. "Document Image Quality Assessment: A Brief Survey." In *2013 12th International Conference on Document Analysis and Recognition*, 723–27. IEEE.