Key Information Extraction from documents: Kleister NDA/Charity challenges

Filip Graliński
filip.gralinski@applica.ai

Applica.ai / Adam Mickiewicz University

March, 15th

Kleister: A novel task for Information Extraction involving Long Documents with Complex Layout

Filip Graliński[1,2], Tomasz Stanisławek[1,4], Anna Wróblewska[1,4],
Dawid Lipiński[1], Agnieszka Kaliska[1,2], Paulina Rosalska[1,3],
Bartosz Topolski[1], Przemysław Biecek[4,5]

[1]Applica.ai, 15 Zajęcza, Warsaw, 00351, Poland, firstname.lastname@applica.ai

[2]Adam Mickiewicz University, 1 Wieniawskiego, Poznan, 61712, Poland, firstname.lastname@amu.edu.pl

[3]Nicolaus Copernicus University, 11 Gagarina, Torun, 87100, Poland, firstname.lastname@umk.pl

[4]Warsaw University of Technology, Koszykowa 75, Warsaw, Poland, firstnameletter.lastname@pw.edu.pl

[5]Samsung R&D Institute Poland, Plac Europejski 1, Warsaw, Poland, firstnameletter.lastname@samsung.com

Abstract

State-of-the-art solutions for Natural Language Processing (NLP) are able to capture a broad range of contexts, like the sentence-level context or document-level context for short documents. But these solutions are still struggling when it comes to longer, real-world documents with the information encoded in the spatial structure of the document, such as page elements like tables, forms, headers, openings or footers; complex page layout or presence of multiple pages.

To encourage progress on deeper and more complex Information Extraction (IE) we introduce a new task (named *Kleister*) with two new datasets. Utilizing both textual and structural layout features, an NLP system must find the most important information, about various types of entities, in long formal documents. We propose *Pipeline* method as a text-only baseline with different Named Entity Recognition architectures (Flair, BERT, RoBERTa). Moreover, we checked the most popular PDF processing tools for text extraction (pdf2djvu, Tesseract and Textract) in order to analyze behavior of IE system in presence of errors introduced by these tools.

1 Introduction

Information Extraction (IE) requires quick but careful skimming through the whole document. We often have to not only search for pieces of information, but also to generate final output for specific entity type (e.g. aggregate multiple occurrences of organization names into one). In practice, this means that the results should be presented in an appropriate form (e.g. data points such as addresses normalized to

Information Extraction Challenges

- Applica Kleister
  - Kleister NDA
  - Kleister Charity
- PolEval 2020 Annual Reports

SPRAWOZDANIE Z DZIAŁALNOŚCI GRUPY KAPITAŁOWEJ ZM „ROPCZYCE" S.A. za I półrocze 2012 r.

1. **INFORMACJE O SPÓŁKACH WCHODZĄCYCH W SKŁAD GRUPY KAPITAŁOWEJ**

JEDNOSTKA DOMINUJĄCA – PREZENTACJA SPÓŁKI

ZM ROPCZYCE S.A.

Zakłady Magnezytowe „ROPCZYCE" S.A. (ZMR S.A.)

Siedziba: **Ropczyce, woj. podkarpackie**
Adres: **ul. Przemysłowa 1, 39-100 Ropczyce**
Regon: **690026060**
NIP: **818-00-02-127**
www.ropczyce.com.pl

PRZEDMIOT DZIAŁALNOŚCI

Przedmiot działalności ZMR S.A. obejmuje produkcję i sprzedaż zasadowych wyrobów og
które są niezbędnym elementem konstrukcji wyłożeń pieców i urządzeń cieplnych pr
wysokich temperaturach, głównie w hutnictwie żelaza i stali, hutnictwie metali nieżelaz
przemyśle cementowo-wapienniczym, odlewniczym.
Spółka świadczy także usługi w zakresie nawęglania i ulepszania cieplnego wyrobów oraz pr
badawczo-rozwojowe w dziedzinie związanej z przedmiotem jej działalności.

`period_from`	2012-01-01
`period_to`	2012-06-30
`postal_code`	39-100
`city`	Ropczyce
...	

`company`, `drawing_date`, `period_from`, `period_to`, `postal_code`, `city`, `street`, `street_no`, `people`


NER?
imgflip.com

. . . no!

This is a Key Information Extraction task, not NER*.

- we are interested in the information not where it is
- not just any person, but CEO, etc.

* But of course you could use NER as a part of the pipeline

KIRKLEES ACTIVE LEISURE

TRUSTEES' REPORT (INCLUDING DIRECTORS' REPORT AND STRATEGIC REPORT)

The Trustees, who are directors of the Charity, present their annual report on the affairs of the Charity and the Group, together with the financial statements and auditor's report for the year ending 31 March 2018. The Trustees have adopted the provisions of the statement of recommended practice (SORP) "Accounting and Reporting by Charities" (FRS 102) in preparing the annual report and financial statements of the Charity.

CHARITY NUMBER
1091226

COMPANY NUMBER
4331165

PRINCIPAL OFFICE
Stadium Business and Leisure Complex,
Sandtun Way,
Huddersfield, HD1 6PG

AUDITORS
Wheawill & Sudworth Limited
35 Westgate
Huddersfield, HD1 1PA

BANKERS
Barclays Bank plc
17 Market Place
Huddersfield, HD1 2AB

STRUCTURE, GOVERNANCE AND MANAGEMENT
Governing Document
Kirklees Active Leisure ("KAL") was formed as a Company Limited by Guarantee, not having share capital and having charitable status, on 29 November 2001. The Charity operates community recreation facilities on behalf of the Local Authority (Kirklees Metropolitan Council) and has one wholly owned trading subsidiary, Kirklees Active Leisure Trading Limited ("KALT"). The Charity is required to comply with both the Companies Act 2006, the Statement of Recommended Practice "Accounting and Reporting by Charities" (FRS 102) and has to meet general Charity Commission regulations.

The Memorandum and Articles of Association are the Charity's constitution.

CHAIRMAN
D. Stephenson

TRUSTEES
J.A. Briggs
M.T. Brooke
S.S. Khela
D.C. Thomson
Cllr. W.J. Dodds (resigned 25/7/18)
A.N. Fletcher
Cllr. M.S. Sokhal
S.M. Sopala
B.C. Stahelin
J.S. Fletcher
D. Morby
Cllr. M.S. Thompson (appointed 25/7/18)

Trustees' Annual Report for the period

From	Period start date Day	Month	Year	To	Period end date Day	Month	Year
	2	Jul	2016		1	Jul	2017

Section A Reference and administration details

CHARITY NAME: The A M Perry Charitable Foundation

Other names charity is known by:

REGISTERED CHARITY NUMBER (IF ANY): 1059073

CHARITY'S PRINCIPAL ADDRESS: NatWest Bank Plc, Trustee Department, 6th Floor, Trinity Quay 2, Avon Street, Bristol, Postcode BS2 0PT

Names of the charity trustees who manage the charity

	Trustee name	Office (if any)	Dates acted if not for whole year	Name of person (or body) entitled to appoint trustee (if any)
1	NatWest Bank Plc			
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

Names of the trustees for the charity, if any, (for example, any custodian trustees)

Name	Dates acted if not for whole year

TAR 1 April 2009

KIRKLEES ACTIVE LEISURE
CONSOLIDATED STATEMENT OF FINANCIAL ACTIVITIES (including Income and Expenditure Account)

for the year ended

31 MARCH 2018

Note		2018 Unrestricted funds £	2018 Restricted Funds £	2018 Total £	2017 Total £
	Income and endowments from:				
2	Charitable activities	12,853,044	85,500	12,938,544	13,280,540
3	Other trading activities	182,573	-	182,573	213,321
4	Investments	5,043	-	5,043	6,407
5	Other	1,583,137	72,442	1,655,579	2,352,285
	Total income	14,623,797	157,942	14,781,739	15,852,553
	Expenditure on:				
6	Raising funds	122,139	-	122,139	169,577
6	Charitable activities	16,196,204	122,508	16,318,712	15,899,868
	Total expenditure	16,318,343	122,508	16,440,851	16,069,445
	Net income/(expenditure)	(1,694,546)	35,434	16,440,851	(216,892)
	Other recognised gains/(losses):				
24	Re-measurement gain/(loss) on defined benefit pension scheme	1,299,000	-	1,299,000	(3,569,000)
	Net movement in funds before tax	(395,546)	35,434	(360,112)	(3,785,892)
	Corporation tax charge on subsidiary	(554)	-	(554)	(226)
21	**Net movement in funds after tax**	(396,100)	35,434	(360,666)	(3,786,118)
	Reconciliation of funds:				
21	Total funds brought forward	(6,879,597)	391,859	(6,487,738)	(2,701,620)
21	**Total funds carried forward**	(7,275,697)	427,293	(6,848,404)	(6,487,738)

All income and expenditure derive from continuing activities.

The statement of financial activities includes all gains and losses recognised during the year.

Charity Name: The Alec Middleton Perry Charitable Foundation — No (if any): 1059073

Receipts and payments accounts — CC16a

For the period from: Period start date 02/07/2016 To Period end date 01/07/2017

Section A Receipts and payments

	Unrestricted funds to the nearest £	Restricted funds to the nearest £	Endowment funds to the nearest £	Total funds to the nearest £	Last year to the nearest £
A1 Receipts					
Income from Securities	21,969	-	-	21,969	20,327
Trustee Deposit Interest	6	-	-	6	45
	-	-		-	
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
Sub total	21,975	-	-	21,975	20,372
A2 Asset and investment sales, etc.	-	-	263,748	263,748	226,388
Total receipts	21,975	-	263,748	285,723	246,760
A3 Payments					
Charitable Payments	24,000	-	-	24,000	17,500
Bank's fees	7,941	-	-	7,941	7,445
Value Added Tax	1,589	-	-	1,589	1,489
Independent Examiner Fee	400	-	-	400	375
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
Sub total	33,930	-	-	33,930	26,809
A4 Asset and investment purchases, etc.	-	-	261,697	261,697	221,917
Total payments	33,930	-	261,697	295,627	248,726
Net of receipts/(payments)	- 11,955	-	2,051	- 9,904	- 1,966
A5 Transfers between funds	9,930	-	- 9,930	-	-
A6 Cash funds last year end	9,986	-	19,072	29,058	31,024
Cash funds this year end					

STATEMENT OF FINANCIAL ACTIVITIES

for the year ended 31 December 2017 incorporating an income and expenditure account

	Notes	Restricted funds (£'000)	Unrestricted funds (£'000)	2017 Total (£'000)	Restricted funds (£'000)	Unrestricted funds (£'000)	2016 Total (£'000)
Income							
Income from:							
– Donations and legacies	2a	16,920	23,297	**40,217**	13,790	27,360	41,150
– Trading	2a	-	18	**18**	-	11	11
– Investment income	2c	-	162	**162**	-	78	78
Income from charitable activities:							
– Grants	2b	11,523	610	**12,133**	18,516	640	19,156
– Trading income from charitable activities	2b	-	2,507	**2,507**	-	11	11
Total Income		28,443	26,594	55,037	32,306	29,100	60,406
Expenditure on:							
Raising funds:							
– Raising funds	4	228	8,895	**9,123**	77	9,240	9,317
– Fundraising trading: costs of goods sold and other costs	4	-	33	**33**	-	5	5
Charitable activities	6	33,450	18,293	**51,743**	30,896	18,402	49,298
Total Expenditure		33,678	27,221	60,899	30,973	27,647	58,620
Net (expenditure) / income		(5,235)	(627)	**(5,862)**	1,333	453	1,786
Net (expenditure) / income for the year before other recognised gains and losses		(5,235)	(627)	**(5,862)**	1,333	453	1,786
Exchange rate gains (losses)		-	(113)	**(113)**	-	569	569
Net movement in funds		(5,235)	(740)	**(5,975)**	1,333	1,022	2,355
Total funds brought forward at 1 January		8,758	9,767	**18,525**	7,425	8,745	16,170
Total funds carried forward at 31 December		3,523	9,027	**12,550**	8,758	9,767	18,525

The notes on pages 65–84 form part of these financial statements. There are no recognised gains and losses

Dataset name	CoNLL 2003	WikiReading	FUNSD	SROIE	Kleister NDA	Kleister Charity
Source	Reuters news	Wikipedia	forms	receipts	EDGAR	UK Charity Com.
Documents	1,393	4.7M	199	973	540	2,778
Pages	—	—	199	973	3,229	61,643
Entities	35,089	18M	9,743	3,892	2,160	21,612
train docs	946	16.03M	149	626	254	1,729
dev docs	216	1.89M	—	—	83	440
test docs	231	0.95M	50	347	203	609
Input/Output on token level	✓	✗	✓	✓	✗	✗
Long Document	✗	✓	✗	✗	✓	✓
Complex layout	✗	✗	✓	✓	✓	✓
OCR	✗	✗	✓	✓	✗	✓

Table: Summary of the existing English datasets and the Kleister sets.

Entities	Description
	NDA dataset
party	parties appearing in the agreement (each of them is treated as a separate entity)
jurisdiction	state or country whose law governs the agreement
effective_date	date on which the contract becomes legally binding
term	duration of the agreement
	Charity dataset
address__post_town	post town (part of a charity address)
address__postcode	postcode (part of a charity address)
address__street_line	street with the house number (part of a charity address)
charity_name	name of the charitable organization
charity_number	identification number in the charity register
report_date	date of reporting
income_annually	annual income in British pounds (GBP)
spending_annually	annual spending in British pounds (GBP)

Evaluation metric

F1-score will be used as the evaluation metric

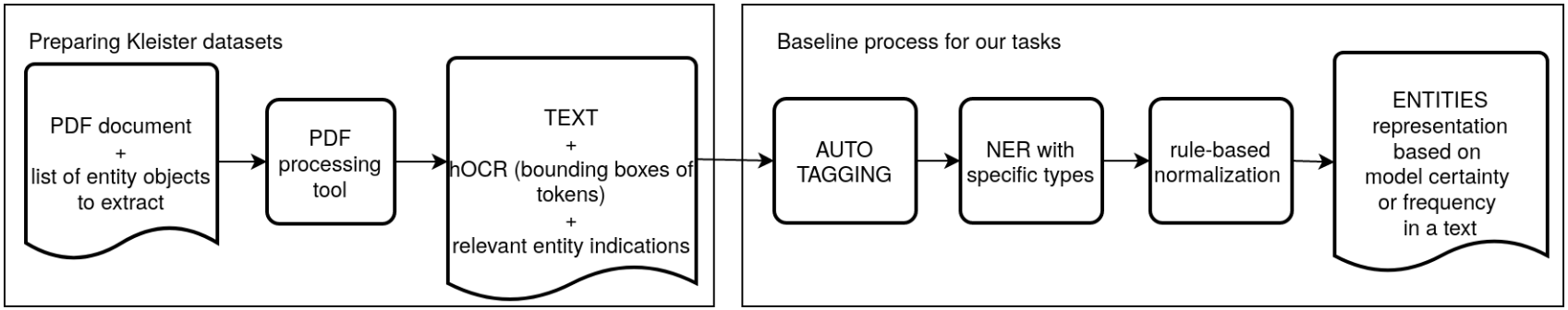
```
--metric MultiLabel-F1
```

Possible approaches

- handcrafted rule, e.g. regexps (as a baseline)
- standard NER (for general entities) + role classification
- specialized NER (but you need to autotag entities)
- end-to-end (generative models)
- ensembles

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Filip Graliński, *LAMBERT: Layout-Aware (Language) Modeling using BERT for information extraction*, `https://arxiv.org/abs/2002.08087`

Autotagging approach

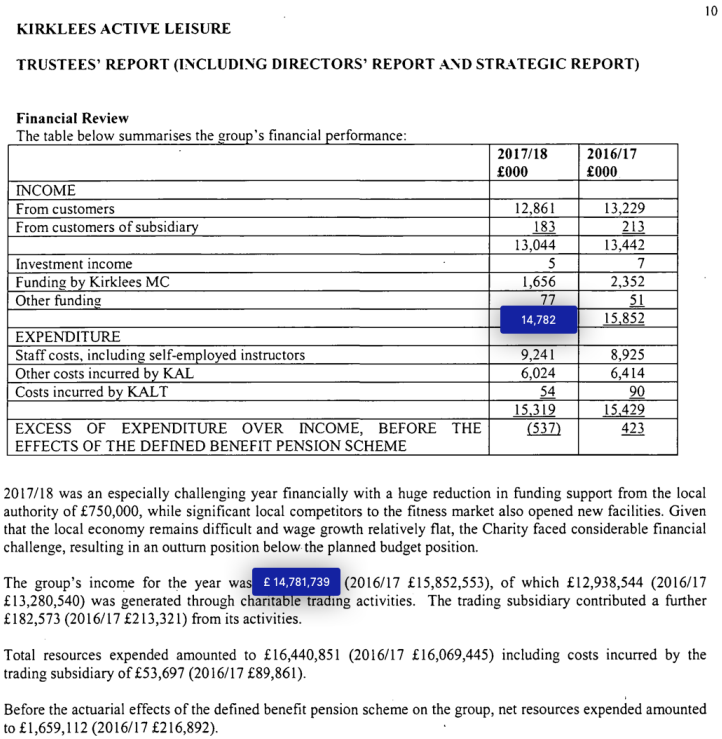


Entity name	**Flair**	**BERT**	**RoBERTa**	**LayoutLM**	**LAMBERT**	Autotag.	Human
Kleister-NDA dataset (pdf2djvu)							
effective_date	79.37	80.20	81.50	82.08	**85.27**	79.00	100 %
party	70.13	71.60	**80.83**	75.28	78.70	33.15	98 %
jurisdiction	93.87	95.00	92.87	94.40	**96.50**	54.10	100 %
term	**60.33**	45.73	52.27	48.34	55.03	74.10	95 %
ALL	77.83	78.20	81.00	78.68	**81.77**	60.09	97.86 %
Kleister-Charity dataset (Azure CV)							
post_town	**83.30**	77.03	77.70	79.97	81.03	66.04	98 %
postcode	82.63	87.10	**88.40**	81.06	82.97	87.60	100 %
street_line	68.17	62.23	72.03	70.92	**75.33**	75.02	96 %
charity_name	72.40	75.93	78.03	78.82	**79.10**	67.00	99 %
charity_number	**96.73**	96.67	95.37	95.76	96.57	98.60	98 %
income	70.93	64.43	69.73	72.86	**76.90**	69.00	97 %
report_date	95.67	96.60	96.77	95.42	**95.80**	89.00	100 %
spending	61.67	67.30	68.60	71.20	**74.33**	73.00	92 %
ALL	80.10	78.33	81.50	80.74	**82.97**	78.16	97.45 %

Comparing OCR engines

Kleister NDA dataset (born-digital PDF files)					
PDF tool	**Flair**	**BERT**	**RoBERTa**	**LayoutLM**	**LAMBERT**
Azure CV	$78.03_{\pm 0.12}$	$77.67_{\pm 0.18}$	$79.33_{\pm 0.68}$	$77.43_{\pm 0.29}$	$80.57_{\pm 0.25}$
pdf2djvu	$77.83_{\pm 0.26}$	$78.20_{\pm 0.17}$	$81.00_{\pm 0.05}$	$78.47_{\pm 0.76}$	$\mathbf{81.77_{\pm 0.09}}$
Tesseract	$76.57_{\pm 0.49}$	$76.60_{\pm 0.30}$	$77.81_{\pm 0.97}$	$77.70_{\pm 0.48}$	$81.03_{\pm 0.23}$
Textract	$77.37_{\pm 0.08}$	$74.83_{\pm 0.45}$	$79.49_{\pm 0.32}$	$77.40_{\pm 0.40}$	$77.37_{\pm 0.08}$
Kleister Charity dataset (mixture of born-digital and scanned PDFs) (*)					
Azure CV	$81.17_{\pm 0.12}$	$78.33_{\pm 0.08}$	$81.50_{\pm 0.23}$	$81.53_{\pm 0.23}$	$\mathbf{83.57_{\pm 0.29}}$
Tesseract	$72.87_{\pm 0.81}$	$71.37_{\pm 1.25}$	$76.23_{\pm 0.15}$	$77.53_{\pm 0.20}$	$81.50_{\pm 0.07}$
Textract	$78.03_{\pm 0.12}$	$73.30_{\pm 0.43}$	$80.08_{\pm 0.15}$	$80.23_{\pm 0.41}$	$82.97_{\pm 0.21}$

Table: F_1-scores for different PDF processing tools and models checked on Kleister challenges test sets over 3 runs with standard deviation. (*) pdf2djvu does not work on scans. We used the Base version of the BERT, RoBERTa, LayoutLM and LAMBERT models.



10

KIRKLEES ACTIVE LEISURE

TRUSTEES' REPORT (INCLUDING DIRECTORS' REPORT AND STRATEGIC REPORT)

Financial Review

The table below summarises the group's financial performance:

	2017/18 £000	2016/17 £000
INCOME		
From customers	12,861	13,229
From customers of subsidiary	183	213
	13,044	13,442
Investment income	5	7
Funding by Kirklees MC	1,656	2,352
Other funding	77	51
	14,782	15,852
EXPENDITURE		
Staff costs, including self-employed instructors	9,241	8,925
Other costs incurred by KAL	6,024	6,414
Costs incurred by KALT	54	90
	15,319	15,429
EXCESS OF EXPENDITURE OVER INCOME, BEFORE THE EFFECTS OF THE DEFINED BENEFIT PENSION SCHEME	(537)	423

2017/18 was an especially challenging year financially with a huge reduction in funding support from the local authority of £750,000, while significant local competitors to the fitness market also opened new facilities. Given that the local economy remains difficult and wage growth relatively flat, the Charity faced considerable financial challenge, resulting in an outturn position below the planned budget position.

The group's income for the year was £ 14,781,739 (2016/17 £15,852,553), of which £12,938,544 (2016/17 £13,280,540) was generated through charitable trading activities. The trading subsidiary contributed a further £182,573 (2016/17 £213,321) from its activities.

Total resources expended amounted to £16,440,851 (2016/17 £16,069,445) including costs incurred by the trading subsidiary of £53,697 (2016/17 £89,861).

Before the actuarial effects of the defined benefit pension scheme on the group, net resources expended amounted to £1,659,112 (2016/17 £216,892).

Figure: Normalization issues for an `income` entity (amount in the table should be multiplied by 1000).

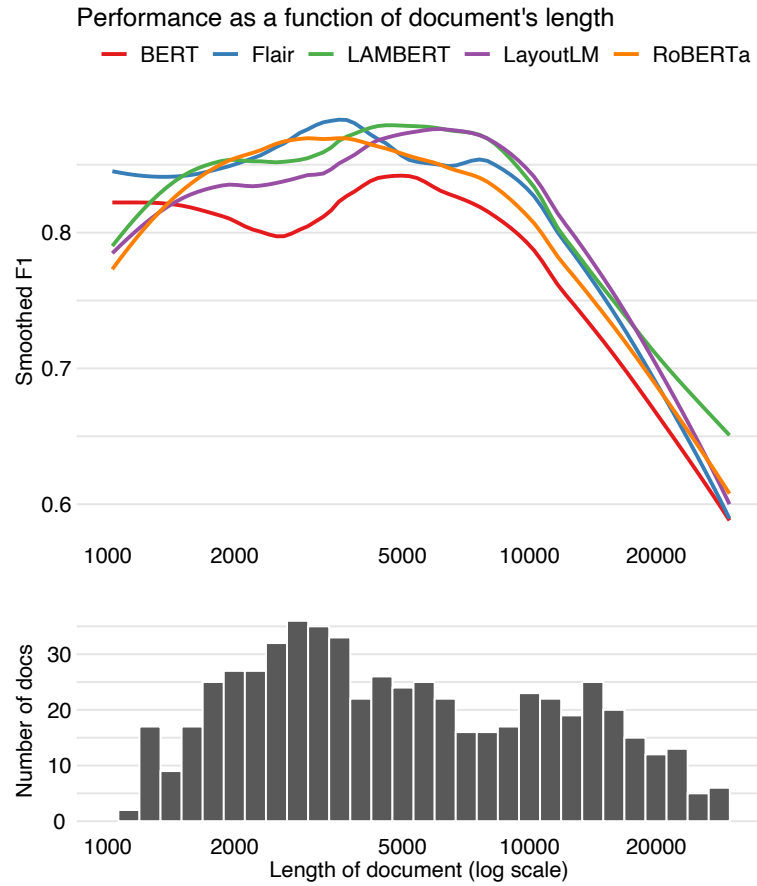


Figure: Relationship between F_1-scores and document length in the Kleister Charity test set for the Azure CV OCR.

Applica Kleister challenges

`https://gitlab.com/kleister-challenge-2021/kleister-nda.git`

`https://gitlab.com/kleister-challenge-2021/kleister-charity.git`

Thank you!