Automatically Inferring the Document Class of a Scientific Article

Antoine Gauquier
antoine.gauquier@ens.fr

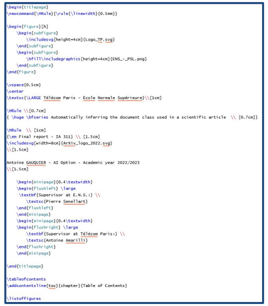
Pierre Senellart
pierre@senellart.com

23rd ACM Symposium on Document Engineering

Session 2 – Document Recognition, Summarisation and Inference
Wednesday, the 23rd of August 2023

Context

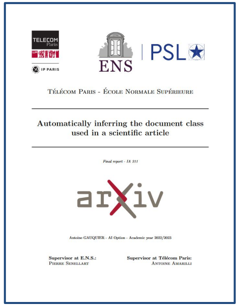
Document class: How the document will be structured and written once in PDF format



```latex
\begin{titlepage}
\newcommand{\HRule}{\rule{\linewidth}{0.5mm}}

\begin{figure}[h]
    \begin{subfigure}
        \includesvg[height=4cm]{Logo_TP.svg}
    \end{subfigure}
    \begin{subfigure}
        \hfill\includegraphics[height=4cm]{ENS_-_PSL.png}
    \end{subfigure}
\end{figure}

\vspace{0.5cm}
\center
\textsc{\LARGE Télécom Paris - Ecole Normale Supérieure}\\[1cm]

\HRule \\[0.7cm]
{ \huge \bfseries Automatically inferring the document class used in a scientific article  \\ [0.7cm]}

\HRule  \\ [1cm]
{\em Final report - IA 311} \\ [1.5cm]
\includesvg[width=8cm]{ArXiv_logo_2022.svg}
\\[1.5cm]

Antoine GAUQUIER - AI Option - Academic year 2022/2023
\\[1.5cm]

    \begin{minipage}{0.4\textwidth}
    \begin{flushleft} \large
     \textbf{Supervisor at E.N.S.:} \\
        \textsc{Pierre Senellart}
    \end{flushleft}
    \end{minipage}
    \begin{minipage}{0.4\textwidth}
    \begin{flushright} \large
        \textbf{Supervisor at Télécom Paris:} \\
        \textsc{Antoine Amarilli}
    \end{flushright}
    \end{minipage}

\end{titlepage}

\tableofcontents
\addcontentsline{toc}{chapter}{Table of Contents}

\listoffigures
```

LaTeX source code

- Scientific style
- Mathematical formulas



Associated document in **PDF** format

Several libraries and commands.

Among them: `\documentclass{report}`

(The argument defines the selected document class)

Association for Computing Machinery (ACM)

`acmart`

American Astronomical Society (AAS)

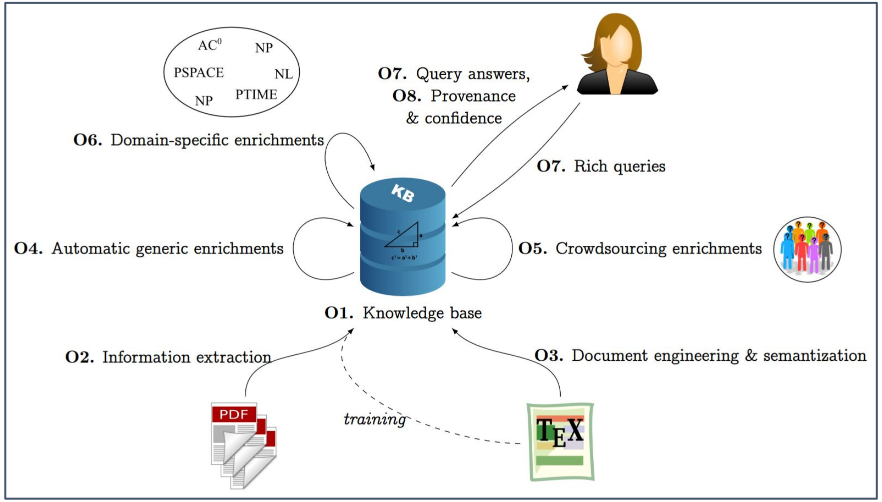
`aastex aastex61 aastex62`

American Mathematical Society (AMS)

`amsart amsproc`

Applications

Systems extracting information from scholarly articles



The TheoremKB project

https://github.com/PierreSenellart/theoremkb

Improve articles indexation in academic search engines

BASE

BASE search engine

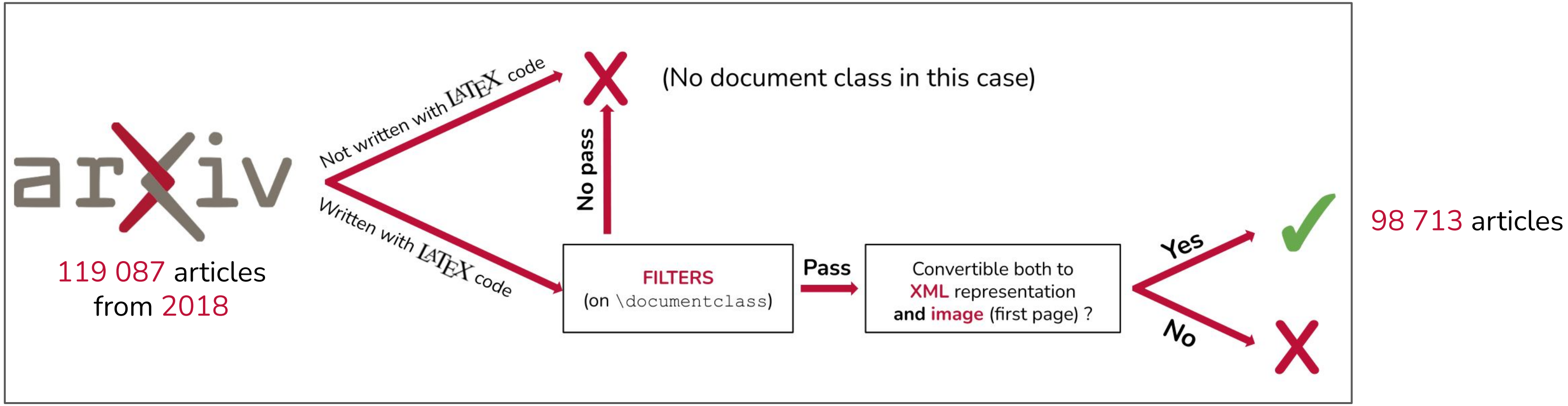
https://www.base-search.net/

Google scholar search engine

https://scholar.google.com/

Outline

Dataset and performance metrics



Among these 98713 articles → more than 1200 document class names. We kept the most frequent ones, and merged the most similar ones (amsart and amsproc for instance), ending in 33 document classes.

$$\text{precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$$

$$\text{recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

$$\text{F}_1\text{-score}_i = 2\frac{\text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i}$$

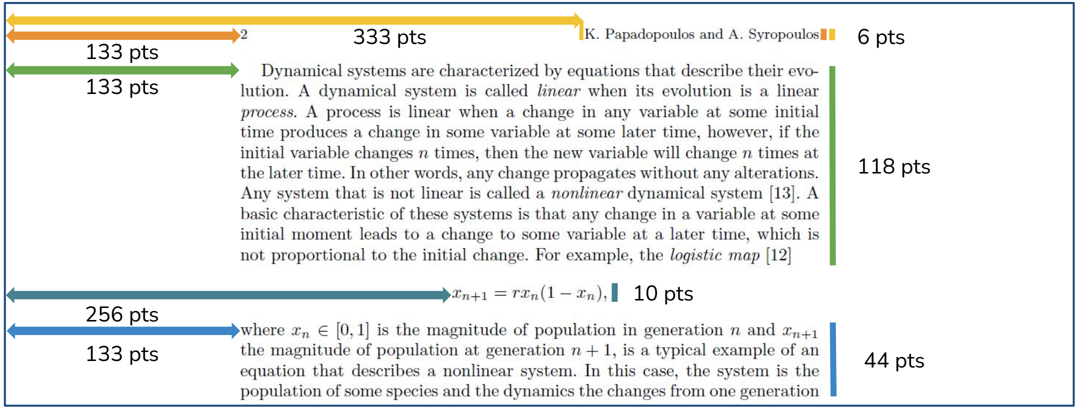
Why **Macroscopic F1-Score** ?

- Macroscopic gives same weight to each document class
- F1-Score gives finer analysis for multiclass classification than accuracy

Statistical study

Construction of five (simple) hand-designed features (1)

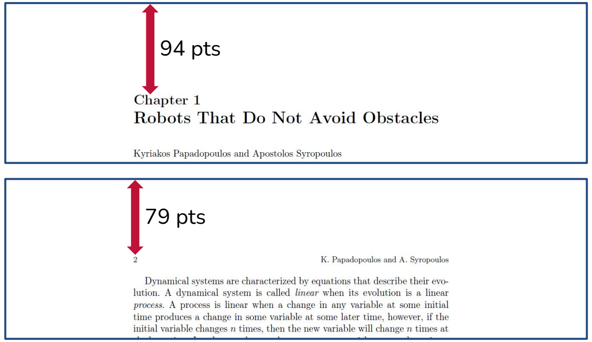
$$1 \text{ point} = \frac{1}{72} \text{ inch}$$



Weighted average left margin **(lm)**

$$\overline{m^{\mathrm{h}}} = \frac{\sum_{i=1}^{N_b} m_i^{\mathrm{h}} \times l_i}{\sum_{i=1}^{N_b} l_i}$$

- m_i^{h} Margin for i-th text block
- l_i Vertical height of i-th text block
- N_b Total number of blocks



Average first top margin **(tm)**

$$\overline{m^{\mathrm{v}}} = \frac{\sum_{i=1}^{N_p} \min_j m_{i,j}^{\mathrm{v}}}{N_p}$$

- $m_{i,j}^{\mathrm{v}}$ Distance between top of page of j-th block of i-th page
- N_p Total number of pages

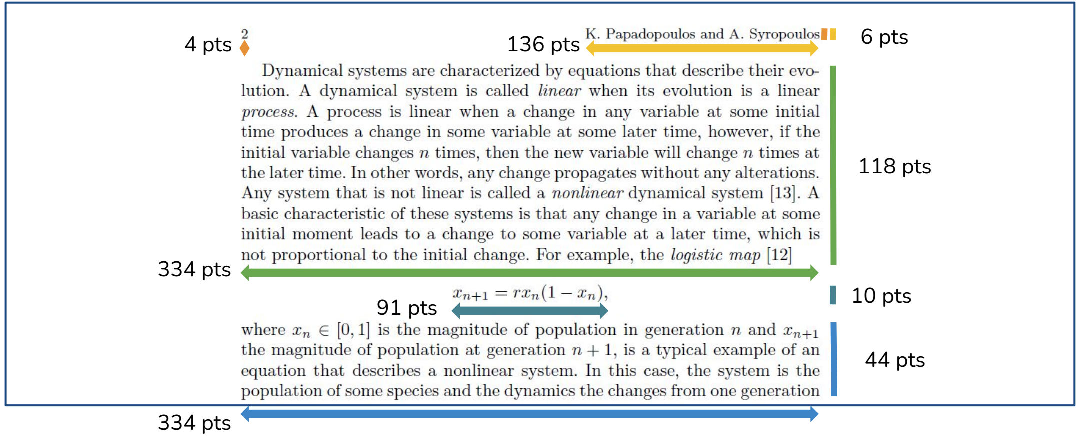
Source: *https://doi.org/10.48550/arXiv.1801.00787*

https://github.com/kermitt2/pdfalto

Statistical study

Construction of five (simple) hand-designed features (2)

$$1 \text{ point} = \frac{1}{72} \text{ inch}$$



2 K. Papadopoulos and A. Syropoulos

Dynamical systems are characterized by equations that describe their evolution. A dynamical system is called *linear* when its evolution is a linear *process*. A process is linear when a change in any variable at some initial time produces a change in some variable at some later time, however, if the initial variable changes n times, then the new variable will change n times at the later time. In other words, any change propagates without any alterations. Any system that is not linear is called a *nonlinear* dynamical system [13]. A basic characteristic of these systems is that any change in a variable at some initial moment leads to a change to some variable at a later time, which is not proportional to the initial change. For example, the *logistic map* [12]

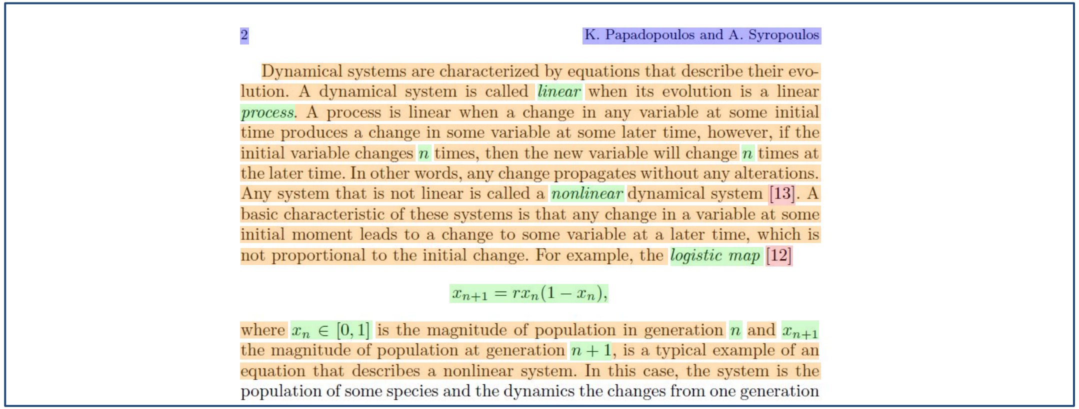
$$x_{n+1} = rx_n(1 - x_n),$$

where $x_n \in [0, 1]$ is the magnitude of population in generation n and x_{n+1} the magnitude of population at generation $n+1$, is a typical example of an equation that describes a nonlinear system. In this case, the system is the population of some species and the dynamics the changes from one generation

Weighted average column width **(cw)**

$$\overline{w} = \frac{\sum_{i=1}^{N_b} w_i \times l_i}{\sum_{i=1}^{N_b} l_i}$$

- w_i Width of i-th text block
- l_i Vertical height of i-th text block



2 K. Papadopoulos and A. Syropoulos

Dynamical systems are characterized by equations that describe their evolution. A dynamical system is called *linear* when its evolution is a linear *process*. A process is linear when a change in any variable at some initial time produces a change in some variable at some later time, however, if the initial variable changes n times, then the new variable will change n times at the later time. In other words, any change propagates without any alterations. Any system that is not linear is called a *nonlinear* dynamical system [13]. A basic characteristic of these systems is that any change in a variable at some initial moment leads to a change to some variable at a later time, which is not proportional to the initial change. For example, the *logistic map* [12]

$$x_{n+1} = rx_n(1 - x_n),$$

where $x_n \in [0, 1]$ is the magnitude of population in generation n and x_{n+1} the magnitude of population at generation $n+1$, is a typical example of an equation that describes a nonlinear system. In this case, the system is the population of some species and the dynamics the changes from one generation

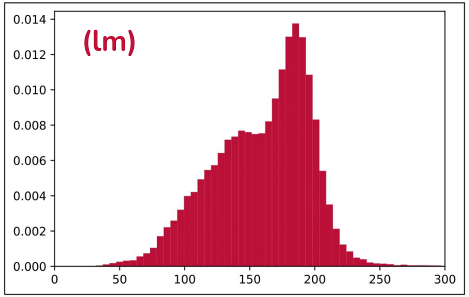
Most common font family **(ff)**
Most common font size **(fs)**

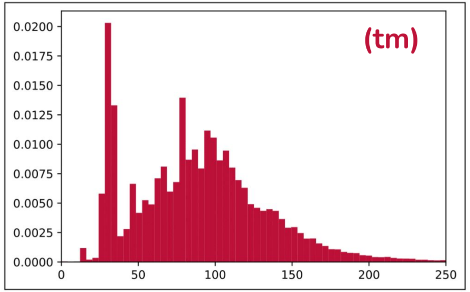
$$f_i = \frac{\sum_{s \in S_i} l_s \times h_s}{\sum_{j=1}^{N_f} \sum_{s \in S_j} l_s \times h_s}$$

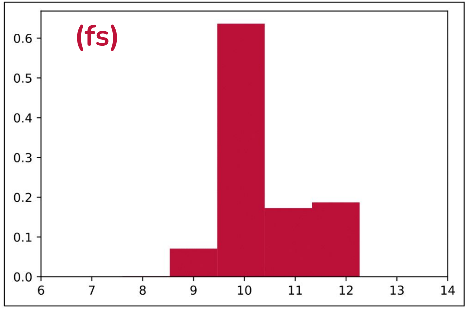
- S_i Set of all tokens of i-th font
- l_s Length of token s
- h_s Height of token s

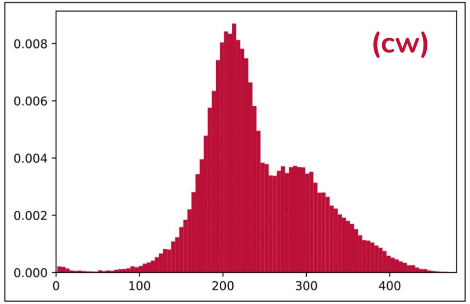
Statistical study

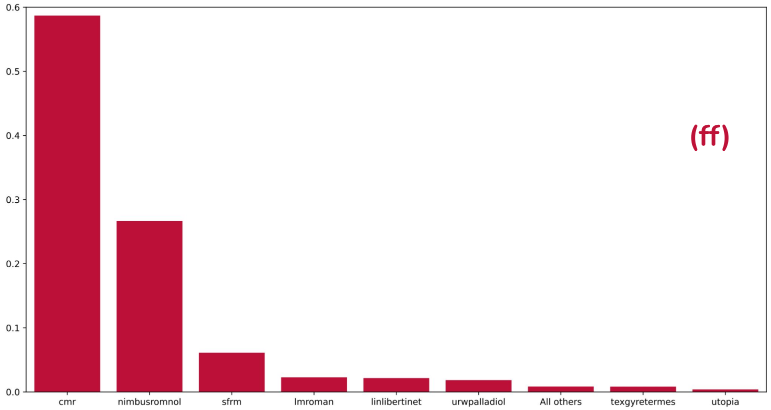
Global distributions of the features











Only a few font families **(ff)** are widely used.

- Gaussian-like distributions for **(lm)**, **(tm)** and **(cw)**.
- Different values for **(fs)** and peak value for **(tm)**.

We identify several characteristics that seem to be document-class specific, and therefore **discriminative**.

Statistical study

Comparison of distributions from two different document classes

Source: *https://doi.org/10.1051/0004-6361/201730392*

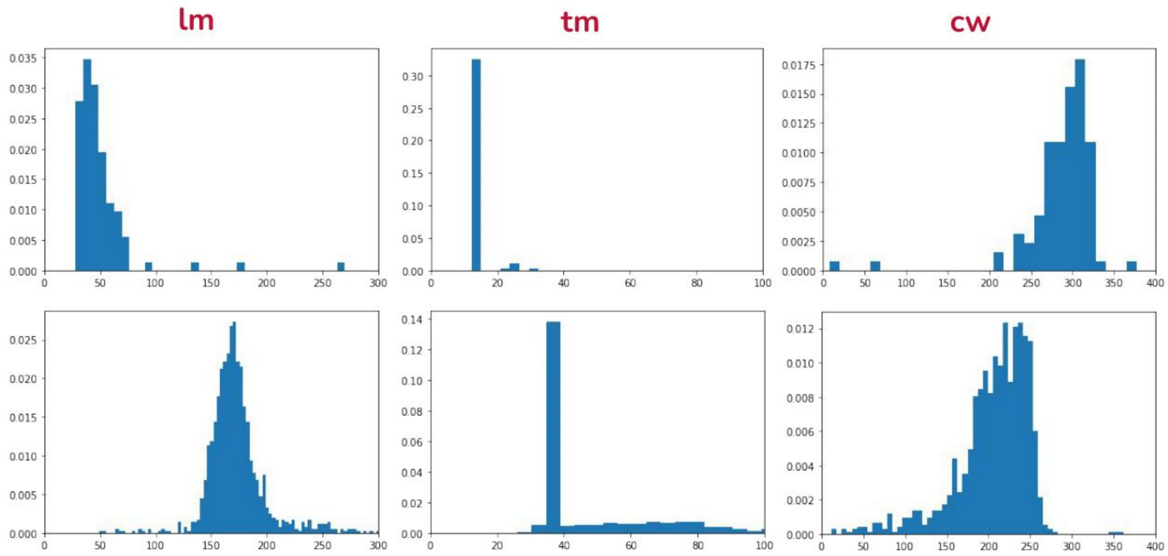
> *SHABANI, SAMADFAM, SADEGHI: LOCAL VISUAL MICROPHONES* 3
>
> the speed of moving hot air or similar transparent fluids against a textured background. In material engineering, Davis *et al.* [9] used sub-pixel motion extraction to estimate certain material properties from a video [9]. In 3D video processing, motion detection is used to extract depth map from binocular images [8, 14].
>
> For sound detection, Zeev *et al.* [25] extracted speech signals by measuring the vibration of people's neck in a video. Davis *et al.* [10] used sub-pixel motion extraction to recover sound from silent video. They extract vibration separately in a number of scales and angles. Then they align signals temporally (to avoid destructive interference) and finally they take a weighted average among all orientations and scales.

bmvc2k document class

Source: *https://doi.org/10.48550/arXiv.1801.09436*

> Mancuso et al.: Long-term evolution of the heliospheric magnetic field intensity
>
> nT (Steinhilber et al. 2010). In the calculation of the 1σ (68.27%) and 2σ (95.45%) confidence intervals for B_{HMF}, we also propagated the uncertainty in τ ($\Delta\tau \approx 0.9$) and the uncertainty in the relation expressing the expected ^{44}Ti production rate Q in a stony meteorite as a function of the modulation parameter ϕ (Eq. 3) estimated as in Usoskin (2006). We remark that the above relation (Eq. 3) has been obtained assuming the LIS of Castagnoli & Lal (1980) in the (North GRIP) ice cores, based on the works of McCracken & Beer (2015) and Usoskin et al. (2015). An average composite time series, $B_{[\mathrm{GCR}]}$, was obtained by averaging together these two independent cosmogenic radionuclide reconstructions in the common intervals from 1766 to 1982 (Fig. 2c).
>
> In Fig. 2, we show the long-term evolution of B_{HMF} obtained in this work along with the above three average

aa document class



This example shows that we can easily **separate** these two document classes with 3 features only.

This entire study indicates that using statistical learning should work pretty well ...

Random forest-based approach

Configuration and results

Random forest model : Ensemble method that uses statistical learning to train a lot of decision trees on different subparts of the training dataset.

Hyperparameters
- Minimum number of samples per leaf → set to 0.01% (risk of overfitting if too high)
- Number of decision trees → set to 1000 to ensure stability of most common decisions

Features of the model : the five hand-designed features. **Output :** Predicted document class among 33 of them.

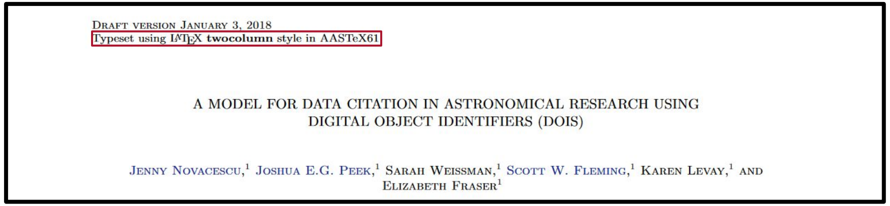
Model	Averaged precision	Averaged recall	Macroscopic F1-Score
Dummy	*0.09%*	*3.03%*	*0.18 %*
Random forest model	64 %	66 %	64 %

Simple modelization (no deep learning and only five, simple, features) → Really promising results !

CNN-based approach

Input data

Text element specific to AAS document class


Draft version January 3, 2018
Typeset using LaTeX **twocolumn** style in AASTeX61

A MODEL FOR DATA CITATION IN ASTRONOMICAL RESEARCH USING DIGITAL OBJECT IDENTIFIERS (DOIS)

Jenny Novacescu,[1] Joshua E.G. Peek,[1] Sarah Weissman,[1] Scott W. Fleming,[1] Karen Levay,[1] and Elizabeth Fraser[1]

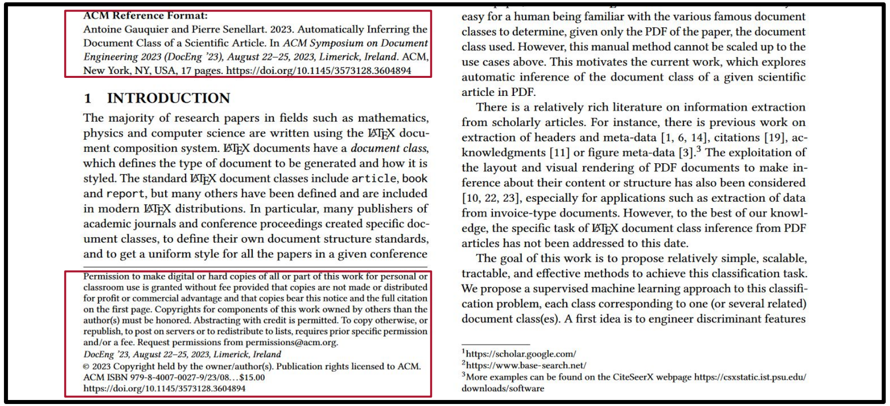
Source : https://arxiv.org/pdf/1801.00004.pdf

Example of input bitmap rendering

256

Source : https://arxiv.org/pdf/1806.06252.pdf

Some usual elements from ACM document class


ACM Reference Format:
Antoine Gauquier and Pierre Senellart. 2023. Automatically Inferring the Document Class of a Scientific Article. In *ACM Symposium on Document Engineering 2023 (DocEng '23), August 22–25, 2023, Limerick, Ireland.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3573128.3604894

1 INTRODUCTION

The majority of research papers in fields such as mathematics, physics and computer science are written using the LaTeX document composition system. LaTeX documents have a *document class*, which defines the type of document to be generated and how it is styled. The standard LaTeX document classes include `article`, `book` and `report`, but many others have been defined and are included in modern LaTeX distributions. In particular, many publishers of academic journals and conference proceedings created specific document classes, to define their own document structure standards, and to get a uniform style for all the papers in a given conference

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DocEng '23, August 22–25, 2023, Limerick, Ireland
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0027-9/23/08…$15.00
https://doi.org/10.1145/3573128.3604894

easy for a human being familiar with the various famous document classes to determine, given only the PDF of the paper, the document class used. However, this manual method cannot be scaled up to the use cases above. This motivates the current work, which explores automatic inference of the document class of a given scientific article in PDF.

There is a relatively rich literature on information extraction from scholarly articles. For instance, there is previous work on extraction of headers and meta-data [1, 6, 14], citations [19], acknowledgments [11] or figure meta-data [3].[3] The exploitation of the layout and visual rendering of PDF documents to make inference about their content or structure has also been considered [10, 22, 23], especially for applications such as extraction of data from invoice-type documents. However, to the best of our knowledge, the specific task of LaTeX document class inference from PDF articles has not been addressed to this date.

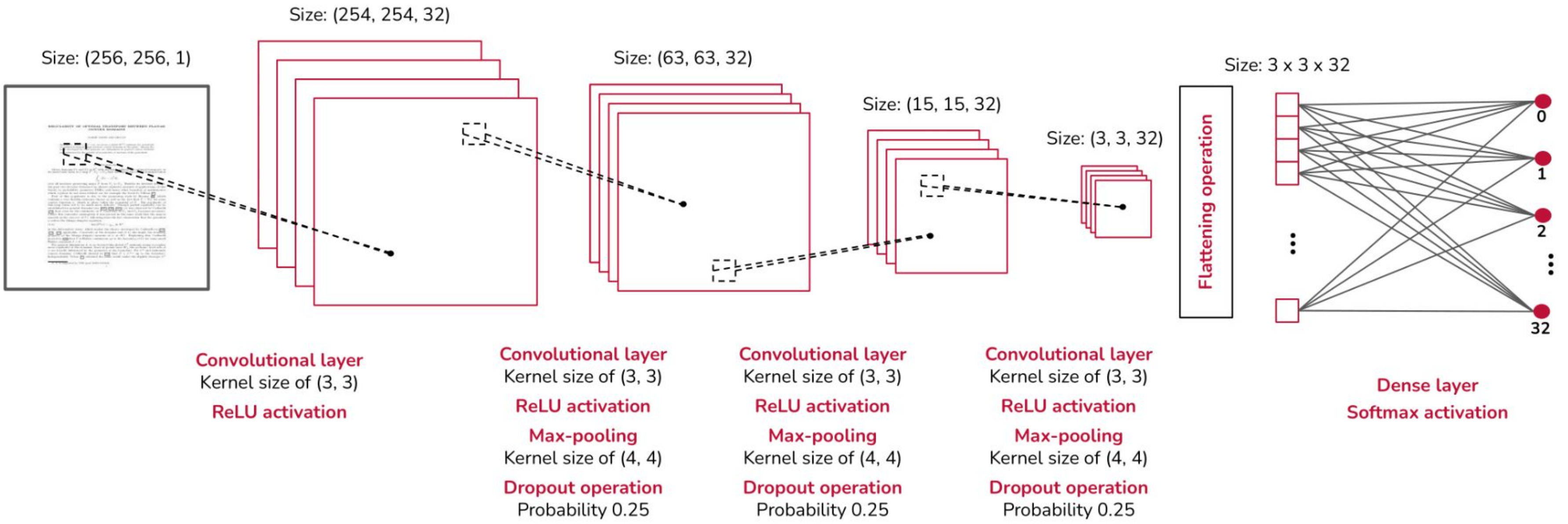
The goal of this work is to propose relatively simple, scalable, tractable, and effective methods to achieve this classification task. We propose a supervised machine learning approach to this classification problem, each class corresponding to one (or several related) document class(es). A first idea is to engineer discriminant features

[1] https://scholar.google.com/
[2] https://www.base-search.net/
[3] More examples can be found on the CiteSeerX webpage https://csxstatic.ist.psu.edu/downloads/software

- ACM Reference Format
- Rights and information about the article

CNN-based approach

Architecture



CNN-based approach

Results and comparison with state-of-the-art

Architecture	Macro F1-Score	Number of parameters	FLOPS (in billions)
Our architecture	*92.31 %*	***38 177***	*1.36*
ResNet50V2	92.28 %	23 632 417	9.13
NASNetMobile	91.31 %	4 304 597	1.50
EfficientNetV2B0	**93.43 %**	4 091 844	**0.80**

Analysis:

- ➔ 100 times less parameters than other models
- ➔ Almost as performant, above 92% of F1-Score
- ➔ Number of floating operations at inference time slightly above EfficientNetV2B0

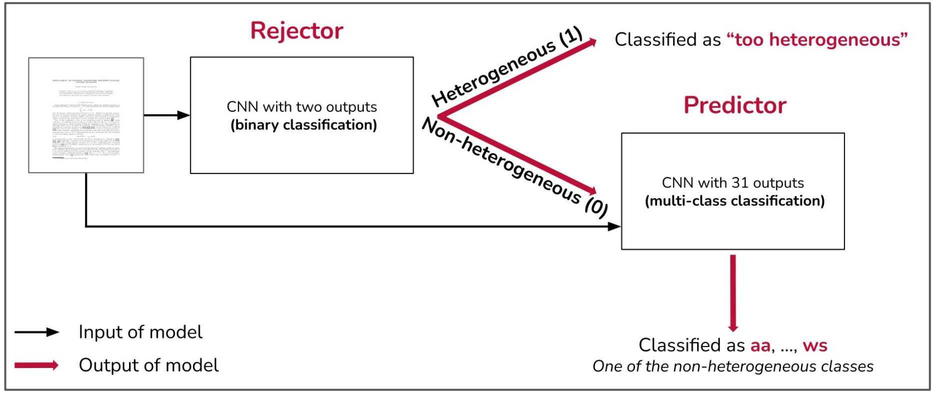
CNN-based approach

Separating heterogeneous document classes with reject option (1)

Document class	Precision	Recall	F1-Score
book	56.84 %	21.39 %	31.09 %
report/wlscirep	52.09 %	77.69 %	62.37 %
other (including article)	69.17 %	65.00 %	67.02 %

Common ground of theses classes : they are widely customizable, and thus embed a great **heterogeneity** of renderings.

What about directly putting apart these heterogeneous classes before applying classifier ? This is **reject option**.



CNN-based approach

Separating heterogeneous document classes with reject option (2)

Model	Precision	Recall	F1-Score
Rejector	90.55 %	89.15 %	89.04 %
Classifier	96.94 %	96.73 %	96.83 %

Improvement in the classifier performance (more than 4.50% in averaged F1-Score). However ...

The rejector has lower performance → overall system not necessarily better !

Still very useful for applications where we know that **heterogeneous classes are not frequently observed or relevant** (for instance, articles from conference proceedings or journals).

Recall for non-heterogeneous class of rejector is above 98 % : non-heterogeneous classes are almost always classified as so.

Conclusion and perspectives

- It is statistically relevant to discriminate document classes on the basis of features from PDF rendering.
- A (relatively) simple classification method on a set of 5 simple features gives promising results.
- Using a computer-vision based approach (CNN) gives really good performance, comparable to state-of-the-art models with way more parameters.
- We can even improve these results by putting apart heterogeneous classes, which are not related to a specific conference or journal.

- The experiment was conducted on a « small » subset of ArXiV (only 2018): what happens on a larger time frame?
- Dependency on ArXiV: we don't know any dataset where document class is readily available.
- We did show that using document class helps detecting mathematical environments (TheoremKB). But finding an efficient way maximising performance is still in progress.

Thank you for your attention!

Any questions?

https://github.com/AntoineGauquier/inferring_document_class_of_scientific_article/

Antoine Gauquier
antoine.gauquier@ens.fr

Pierre Senellart
pierre@senellart.com