Structure of Scientific Articles

Salma Kastali
salma01@ads.uni-passau.de

Dzejlana Karajic
karaji01@ads.uni-passau.de

Zied Dammak
dammak01@ads.uni-passau.de

ABSTRACT

This paper presents a pipeline for the identification of the structure of scientific papers. Large journals share explicitly their guidelines so participants know exactly how to structure their publications but in most of the medium size conference these guidelines are not explicitly described so participants usually go with the way previous work has been presented.
The idea of this paper is to build a pipeline to retrieve effective and valid information from scientific articles belonging to different small size conferences. And our goal is to find their guidelines. The process of identification of the structure of scientific articles in this work starts with data acquisition, data preprocessing, feature extraction and finally evaluation of the features.
A statistical analysis over the quantitative features shows some similarity across conferences and also identifies the features that distinguish them from each other. At the end, the paper uses a classification algorithm that aims to evaluate the importance of the extracted features in the classification of papers in the right conferences which give us the most important characteristics of a conference.

KEYWORDS

Computational science, Document structure analysis, Regular expressions

INTRODUCTION

Scientific articles get rejected for several reasons; inaccurate studies, irrelevance of the content to the submission target or the weak research motive[5]. Rejected articles can also result from not following the submission requirements, for example minimum number of pages or paper structure. Therefore, large journals such as ACM, explicitly share their submission guidelines[1] that have to be fulfilled in order for a paper to be considered for acceptance. In contrast, small conferences usually do not have the predefined set of guidelines so the articles are written conventionally based on previous work.

Objective. We aim to understand if small to medium sized conferences have common structure by doing a comparison study of the validity of the features for each conference and highlighting the most important ones.

Contribution. We contribute a study exploring the common structure of articles from different small conferences in order to generate a set of requirements for the general structure of an article. In this project we create an end-to-end pipeline for structure analysis. It contains an effortless process for data collection and parsing. In addition, we provide an undemanding and flexible process for data cleaning and feature extraction, and output ready for many statistical models.

Outline. This paper is organized as follows: Section 1 provides an overview of related work. Section 2 details research design and posed research questions. In Section 4 we present our data analysis workflow while in Section 6 we discuss our findings before concluding the paper in Section 7.

Computational Social Science Lab, 2021, Uni Passau
.

1 RELATED WORK

Few works have addressed the identification of the structure of scientific articles. EGLIN and Bres [4] based their research on the visual layout of the document. They explored text and non text areas by extracting the entropy, visibility and geometrical characteristics in order to classify each paper.
Thanh Dien et al. [3] introduced a natural language processing approach for article characterization based on the topic. After using the appropriate text vectorization technique (TFIDF) they applied SVM, Naive Bayes and KNN as classification methods.
Another instance is a large scale analysis, where Boyack et al. analyze the in-text citations on over 5 million articles [2]. The study focuses on interval of citation and citation counts of references and shows that there are field-level differences within them.
In this work we aim to create a computational identification of the scientific articles. Our main focus is to find those similarities inside specific groups of papers and specifically in the computer science discipline.

2 RESEARCH DESIGN

We present the research design including research questions posed in this study. Figure 1 provides an overview of the overall research method, which we explain in detail in subsequent sections.

2.1 Research Objective & Research Questions

The overall objective of this project is to analyze the structure of scientific documents coming from different conferences. The goal is to find specific guidelines for each conference and if a common structure can be defined.
In particular, we analyze five small to medium conferences in computer science with 30 paper each. For this, we pose the research questions presented in Table 1.

3 PIPELINE

Our data pipeline consists of three main phases: *data acquisition, data preprocessing and evaluation phase.* In Figure 2 we display main set of processes, from obtaining raw data to evaluating results. Furthermore, we explain these in detail in next sections.

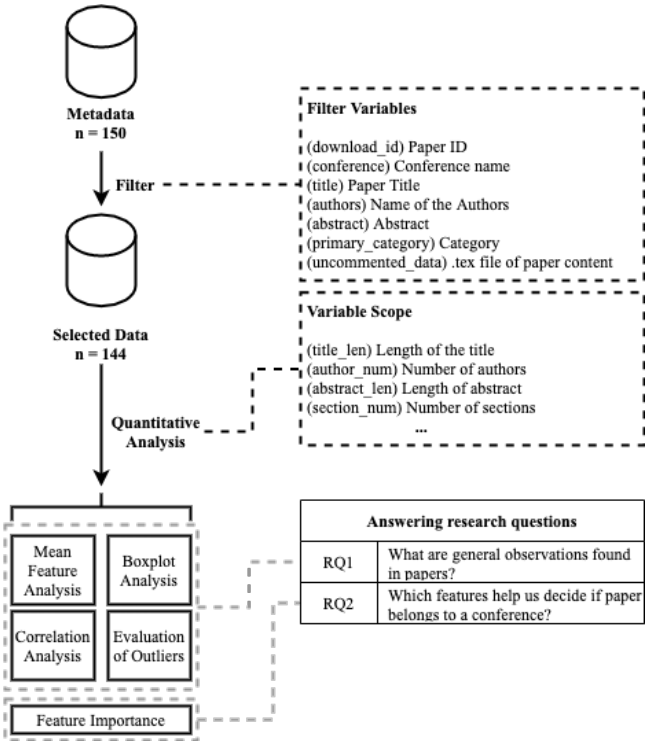


Figure 1: Overview of the research method including the data (variables) selected for the study

Table 1: Overview of the research questions of this study

	Research Question and Rationale
RQ1	*What are the general observations we can find in papers?* In the first step, we start with quantitative data analysis after extracting relevant features and perform statistical methods listed in Figure 1. In this study, we are primarily interested in identifying candidate features and, therefore, we analyze those descriptively only.
RQ2	*Which features help us decide if paper belongs to a conference?* In the second step, we perform feature importance on variables identified in RQ1. In more detail, we build a Classification Model uncovering the features that help classify an article to the right conference.

4 DATA ACQUISITION

Data used in this study was obtained directly utilizing the arXiv API[1] that generates a CSV file containing all available metadata regarding papers. Consequently, the collected raw data was filtered and cleaned in preprocessing step as can be seen in Figure 2. Details and specifics are explained in the following subsections.

[1]https://arxiv.org/help/api

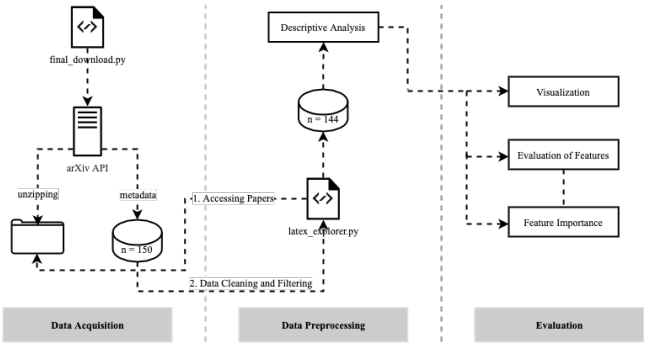


Figure 2: Overview of the data pipeline model used in the study

As previously mentioned, this study collected data using arXiv API, and bellow we present steps and measures in the data acquisition procedure.

i **Defining a search query**
arXiv website was queried in the 'Journal Reference' field for the following conferences: *ICML, ICLR, IJCAI, ECAI and ECML PKDD*. Querying in this specific field assures that papers obtained are within the search domain and maximum obtainable results were set to 200 with 3 iterations so that as many as possible papers get selected. The query returned a CSV file of total 150 papers and the available metadata for each.

ii **Excluding irrelevant papers**
To prepare for the data analysis and avoid downloading unnecessary papers, in this step we inspected and cleaned the data. Specifically, we analyzed the data for NA or missing values for the 'Journal Reference' field, in which case we cannot link a paper to the specific conference. Furthermore, we drop any entry in the same field that contains a term, workshop, ensuring our final results are only papers accepted or presented at the conference/journal.

iii **Avoiding duplicates**
We devise a strategy to collect the IDs of the papers in a set of so-called *good keys* and as we run more than one iteration, this strategy ensures we keep only unique IDs and incrementally add non-existing ones.

iv **Downloading the paper source**
Straightforward process in which our download function takes the list of IDs from previous step and downloads the source files using arXiv API. The output is a compressed list of folders for each conference.

v **Unzipping files**
In this final step, we utilize an unzipping mechanism that loops through the compressed files and returns us uncompressed folders that are easily accessible.

5 DATA PREPROCESSING

The data acquisition phase has generated 2 outputs: the meta_data CSV file and the repository of the conferences stored on the local device. The goal of the preprocessing phase is to access individual

papers of every conference and extract the paper content to the field in the meta_data dataset. Bellow we present the steps and measures in the data preprocessing phase.

i **Files access**
We use a recursive mechanism to loop through the files in each repository. We come across three different cases:
(a) One .tex file: indicating that the article content can be found in the exactly one file that is ready for the data cleaning and feature extraction.
(b) Multiple .tex files: indicating that the article content can be found in the more than one file. A merging step is necessary.
(c) No .tex files: either .pdf format or the file has wrong extension. The paper is dropped.

ii **Merging tex files**
This merging step is launched if the case 2 in the previous step is detected, meaning we have multiple .tex files. In this case, we parse the main .tex file for all *"include"* commands and we change them with the actual content from the other .tex files using a recursive function .

iii **Removing commented lines**
After having our data in only one place for each paper, we delete the commented statements and we add this data in a a new column in our meta_data dataset.

iv **Feature Extraction**
We use Regex expression on the uncommented data to extract qualitative data features like the Title , the section , the subsection and the figures. The first 2 columns of the Table 4 shows the qualitative feature extracted.
From the sections columns, we extracted 5 others dummy columns that represent the existence of a certain section name . The heatmap of the Figure 3 shows the first 30 papers in the y-axis and the sections' name in the x-axis. The idea here is that the white values represent the absence of that section name in a paper. For example the first column "0" shows that 5 papers do not use the section name "Introduction" and the 5th column "4" shows that most of the 30 papers do not use the section name "Acknowledgments".

5.1 Quantitative Analysis

Quantitative analysis consists of multiple consequent steps from variable selection to feature extraction. In paragraphs bellow we give a detailed description on the performed processes.

5.1.1 Variable Selection. In this paper, we focus on the variables that we can quantitatively analyse and compare. Therefore, we only consider those from which we can extract interesting features. This selection defines the base dataset for our analysis, which can be seen in Figure 4. These include download_id, conference, title, authors, abstract, primary_category, uncommented_data as shown in Figure 1. The rest of the variables in meta_data dataset are excluded since they are out of the scope of this study.

5.1.2 Data Selection & Cleaning. Accessing and downloading papers from arXiv yielded 150 entries for 5 different conferences, each having exactly 30 articles. In order to proceed with feature extraction step it was necessary to filter the following variable;

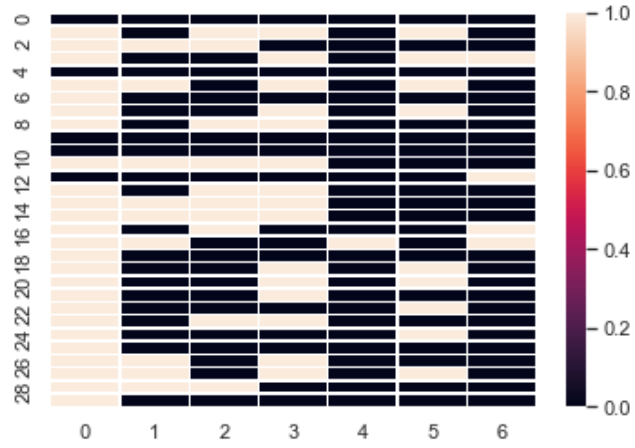


Figure 3: 'Introduction', 'Related Work', 'Experiments', 'Conclusion', 'Acknowledgments','Background' and 'Discussion' Heatmap for 30 papers

	download_id	conference	title	authors	summary	primary_category	uncommented_data
0	0907.0809v1	ICML	Learning as Search Optimization: Approximate L...	[arxiv.Result.Author('Hal Daumé III'), arxiv.R...	Mappings to structured output spaces (strings,...	cs.LG	\documentclass{article}\n\usepackage[accepted]...
1	1305.1704v1	ICML	The Extended Parameter Filter	[arxiv.Result.Author('Yusuf Erol'), arxiv.Resu...	The parameters of temporal models, such as dyn...	stat.ML	\n\documentclass{article}\n\n\usepackage{tight...
2	1602.01783v2	ICML	Asynchronous Methods for Deep Reinforcement Le...	[arxiv.Result.Author('Volodymyr Mnih'), arxiv....	We propose a conceptually simple and lightweig...	cs.LG	\documentclass{article} \usepackage{times}\n\u...
3	1603.03629v2	ICML	Square Root Graphical Models: Multivariate Gen...	[arxiv.Result.Author('David I. Inouye'), arxiv...	We develop Square Root Graphical Models (SQRt)....	stat.ML	\documentclass{article}\n\pdfoutput=1\n\usepac...
4	1609.00288v2	ICML	A Unified View of Multi-Label Performance Meas...	[arxiv.Result.Author('Xi-Zhu Wu'), arxiv.Resul...	Multi-label classification deals with the prob...	cs.LG	\n\documentclass{article}\n\n\usepackage{times...

Figure 4: Snapshot of the first five entries of the base dataset: *'descriptive_ds'*

Table 2: Distribution of papers per conference in the base dataset

ICML	ICLR	ECAI	IJCAI	ECML PKDD
30	30	25	29	30

'uncommented_data', for missing values since it contains article content from which features like number of sections or citations will be extracted. After filtering the dataset the number of entries reduced from 150 to 143. Isolating the 6 entries after filtering we found out that 5 were PDF files so we had to drop them, since our study focuses on specifically analysing latex format of the papers. The remaining entry was downloaded without a proper extension which was manually fixed. Total count of papers per conference is shown in Table 2.

5.1.3 Feature Extraction. This part of the quantitative analysis deals with extracting new and relevant features using Python package for regular expressions[2]. For this study, the usage of regular expressions was a sufficient way to extract relevant data, since the

[2]https://docs.python.org/3/library/re.html

data collection procedure itself focuses on latex format of the articles. This allows us to know from exactly where to obtain needed data knowing Latex commands, their syntax and bounds. Finally, there are two approaches used during feature extraction:

(1) **Extracting new features from existing variables.** In this approach the length of *title, authors and abstract* is calculated to give us new quantitative features; *title_len, authors_num and abstract_len.* Feature description can be seen in Table 3.
(2) **Extracting new features from 'uncommented_data'** using different regular expression patterns which can be seen in Table 4.

Table 3: New features extracted from existing features

Existing feature	New Feature	Description
title	title_len	Character length of the title
authors	author_num	Total number of authors
abstract	abstract_len	Character length of the abstract

6 EVALUATION

The evaluation of our features is by means of the visualisations and looking at the impact of the features on a classification model.

6.1 Feature Analysis

Figures and statistics will provide hard facts about the quality of our features and help us figure out the specific characteristics of each conference.

i **Bar plots**
The bar chart is useful when it comes to exploring and understanding our data points distribution to perform a comparison across different conferences.
The Figure 5 plots our features on one chart axis, and values on the other axis. Each categorical value claims a group of bars, and the length of each bar corresponds to the bar's value. Bars are plotted on a common baseline to allow for easy comparison of values.
We can see which features present highest value, and how other conferences compare against the others.

ii **Correlation Matrix**
In the Figure 6, correlation coefficients is colored according to the value which denotes the strength of the relationship between tow variables.
A high, positive correlation values indicates that the variables measure the same characteristic. If the items are not highly correlated, then the items may measure different characteristics or may not be clearly defined.
A positive linear relationship exists between "Figures" and "Captions", "Sections" and "Figures". This indicates that there is a moderate positive relationship between these variables.
A negative linear relationship exists for "Sections" and "Title length" with negative Pearson correlation coefficients which indicates that, as the number of Sections increases, the title length decrease, and vice versa.

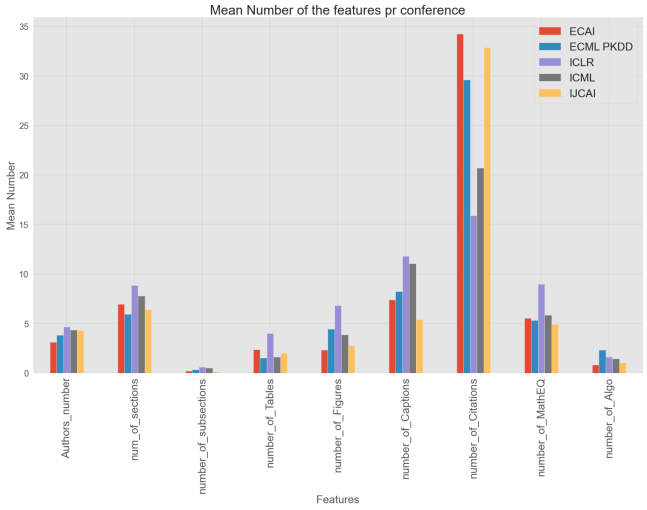


Figure 5: Bar chart

Figure 6: Correlation Matrix

iii **Evaluation of outliers**
We wanted to look at the distribution of features for each conference and see if we can detect some extreme values. The Figure 7 shows that for the conference "ICLR", there is a paper with a number of figures higher than the norm. This lead us to evaluate the histograms of each feature for all conferences.
The Figure 8 shows the frequency distributions for each feature for the "ICLR" conference. The histogram summarizes

Table 4: New features extracted using regular expressions

Existing feature	Regex	New feature	Description
sections	`r'\\section{.*?}`	section_num	Total number of sections
subsections	`r'..\\subsection{.*?}'`	subsection_num	Total number of subsections
tables	`r'begin{table}'`	table_num	Total number of tables
figures	`r'begin{figure}'`	figure_num	Total number of figures
captions	`r'\\caption{.*?}'`	caption_num	Total number of captions
citations	`r'\\cite{.*?}'`	citation_len	Total number of citations
mathEQ	`r'begin{equation}'`	mathEQ_num	Total number of math equations
algo	`r'begin{algorithm}'`	algo_num	Total number of algorithms

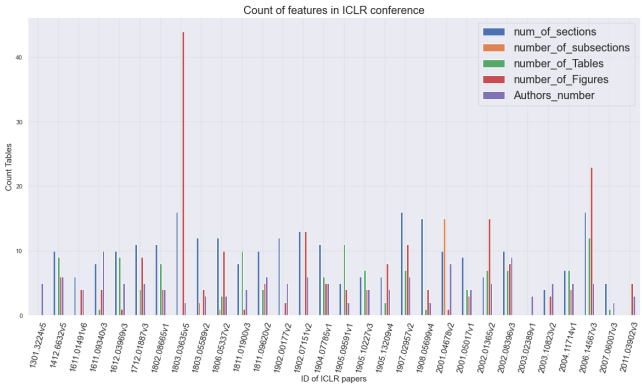


Figure 7: Evaluation of distribution of features for the ICLR conference

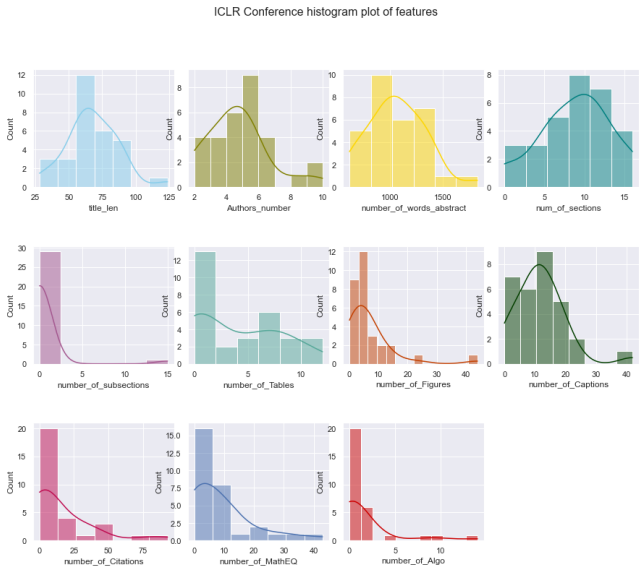


Figure 8: Histograms for ICLR conference

our discrete data by showing the number of data points that fall within a specified range of values and the ones that fall far from the distribution that represents the extreme values.

iv **Most common sections**

An evaluation of the most common section is crucial as it gives us an idea how the conferences chose to start and end their papers.So a count of the frequency of the section name is evaluated in this part. The problem here is that not all the papers choose the same terminology for example : "Conclusion", "Conclusions", "Conclusion and Discussion" .
Wordcloud is useful for quickly perceiving the most prominent sections and for locating a term alphabetically to determine its relative prominence. The Figure 9 shows the most common sections in the ICLR conference.

Figure 9: Word cloud for Sections for the ICLR conference

v **Box Plots**

Box plots provide a visual summary of the data enabling us to quickly identify mean values, the dispersion of the data set, and signs of skewness. The notched boxplot in the Figure 10 shows the division of the data into sections that each contain approximately 25% of the data in that set.

The notched part that allows us to evaluate confidence intervals for the medians of each boxplot by narrowing of the box around the median. Notches are useful in offering a rough

guide to significance of difference of medians; for example if the notches of two boxes do not overlap, this offers evidence of a statistically significant difference between the medians.

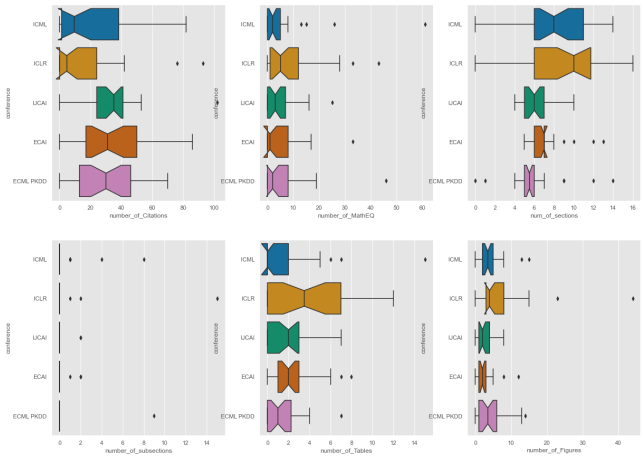


Figure 10: Box Plots

6.2 Feature importance

In this section we test some classification models on our data with the purpose of ranking the features from the most to the least important. We tested Decision Tree classifier and a Random Fores tClassifier model that gave respectively an accuracy score of 34% and 39%. Even though these scores are low , our purpose is only to evaluate the feature importance from these classifiers.
The Feature importance based on a Random Forest Classifier showed the features that best help the model decide if paper belong to a conference. After Create training and test split and Train the model using Random Forest Classifier. The Classifier collects the feature importance values after fitting the model.
The outcome of feature importance stage is a set of features along with the measure of their importance. Once the importance of features get determined, the features can be selected appropriately.

6.3 Results

We extracted features from 30 latex papers coming from 5 conferences. The main remarks we can see is that:

(1) Some conferences share common distributions when considering some features (e.g: Math equations number for ECAI and ECML)
(2) Some conferences can be distinguished based on some features (e.g: ECAI tends to have more citations the ICLR)

We can conclude that we can distinguish conference from one anothers but in some features only. This means that we can define guidelines based on the mean values of the the quantitative data. But these features do not give us enough information to be able to classify a new incoming paper to the right conference. Further feature engineering techniques would help having better results.

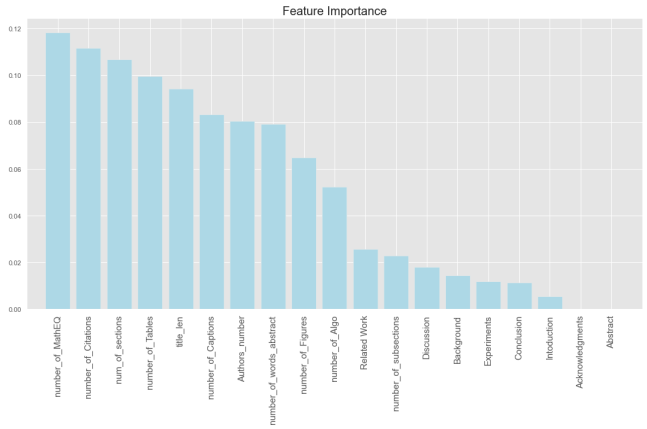


Figure 11: Feature importance using Random Forest Classifier

7 DISCUSSION & CONCLUSION

Our features do not give us enough information to be able to classify papers in right conference. For production-level models, we need to train on datasets larger than 5 Conferences with more than 100 papers each. We can also add an analysis of the PDFs to extra information regarding the layout, font, size and position of the features.

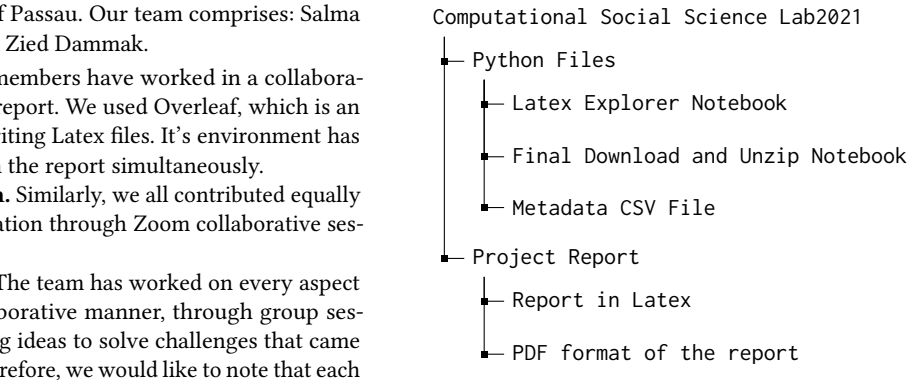
8 ACKNOWLEDGEMENT

This article was written during the Computation Social Science Lab 2021 at the University of Passau. Our team comprises: Salma Kastali, Dzejlana Karajic and Zied Dammak.

- **Project Report.** All members have worked in a collaborative way writing this report. We used Overleaf, which is an online platform for writing Latex files. It's environment has allowed us to work on the report simultaneously.
- **Project Presentation.** Similarly, we all contributed equally preparing the presentation through Zoom collaborative sessions.
- **Jupyter Notebooks** The team has worked on every aspect of the code in a collaborative manner, through group sessions by brainstorming ideas to solve challenges that came across each phase. Therefore, we would like to note that each member of this team contributed equally to this project and it wouldn't be fair to any to assign them a part of this work, since it is truly a teamwork effort.

A FOLDER STRUCTURE

Bellow we present all files that we are submitting with this report.



REFERENCES

[1] [n.d.]. Submitting Articles to ACM Journals. https://www.acm.org/publications/authors/submissions-top

[2] Kevin W. Boyack, Nees Jan van Eck, Giovanni Colavizza, and Ludo Waltman. 2018. Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics* 12, 1 (2018), 59–73. https://doi.org/10.1016/j.joi.2017.11.005

[3] Tran Thanh Dien, Bui Huu Loc, and Nguyen Thai-Nghe. 2019. Article Classification using Natural Language Processing and Machine Learning. In *2019 International Conference on Advanced Computing and Applications (ACOMP)*. 78–84. https://doi.org/10.1109/ACOMP.2019.00019

[4] Véronique Eglin and Stéphane Bres. 2003. Document page similarity based on layout visual saliency: Application to query by example and document classification. 1208–1212. https://doi.org/10.1109/ICDAR.2003.1227849

[5] Suvarna Khadilkar. 2018. Rejection Blues: Why Do Research Papers Get Rejected? *The Journal of Obstetrics and Gynecology of India* 68 (07 2018). https://doi.org/10.1007/s13224-018-1153-1