Experiment Design for Data Sciene: Exercise 2

Paper Reproduction: Option 1

Konstantin Damnjanovic
01151948

Moritz Leidinger
11722966

Dzan Operta
11935976

ABSTRACT

In this project we are trying to reproduce the results, mainly the three tables featured, of the paper *TUD-MMC at MediaEval 2016: Context of Experience task* [1], which is on the other using a data set described by the paper *Right inflight? A dataset for exploring the automatic prediction of movies suitable for a watching situation. In Proceedings of the 7th International Conference on Multimedia Systems* [2]

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Konstantin Damnjanovic, Moritz Leidinger, and Dzan Operta. 2018. Experiment Design for Data Sciene: Exercise 2: Paper Reproduction: Option 1. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 DATA PREPROCESSING

The paper worked with audio data, text data, visual data, metadata and user rating. The training and test data split was already done, as the corresponding files were in separate folder. However, finding the corresponding target values, the binary variable if the movie is good to watch on an airplane or not, was not as straight forward. While the spreadsheet that listed all movies that were in the dev set, contained the column 'goodforairplane', the file for the test set did not. In the whole folder structure there were some spreadsheets labeled 'test set' that contained the target value, but none of these had a complete match with the data files in the folders. We ended up merging the test data file with the 'dataset_complete.xslx', in the tab 'test', from the mediaeval folder. The detail on the steps on data processing that was done really varied, so we will be discussing the different types of data separately in more detail.

1.1 Audio Data

The audio data was comprised of one file per movie, each having 13 rows. The number of columns changed from movie to movie, but was constant for each file. We assumed the length was some kind of measure per frame or other increment of the triler. This one was probably the most well described part of the data sets inthe paper, so the preprocessing was pretty staright forward. They mention each row represents a Mel-frequency cepstral coefficient (MFCC) and that they took the average per row, filling any missing value with 0. We did just that and formed 13 audio features per movie.

Table 1 from the paper

Features used	Precision	Recall	F1
User rating	0.371	0.609	0.461
Visual	0.447	0.476	0.458
Metadata	0.524	0.516	0.519
Metadata + user rating	0.581	0.6	0.583
Metadata + visual	0.584	0.6	0.586

1.2 Visual Data

The visual data was not as transparent.

1.3 Textual Data

1.4 Metadata & User Scores

2 REPRODUCING TABLE 1

As noted in the introduction, the "`acmart`" document class can be used to prepare many different kinds of documentation — a double-blind initial submission of a full-length technical paper, a two-page SIGGRAPH Emerging Technologies abstract, a "camera-ready" journal article, a SIGCHI Extended Abstract, and more — all by selecting the appropriate *template style* and *template parameters*.

This document will explain the major features of the document class. For further information, the *LaTeX User's Guide* is available from https://www.acm.org/publications/proceedings-template.

The primary parameter given to the "`acmart`" document class is the *template style* which corresponds to the kind of publication or SIG publishing the work. This parameter is enclosed in square brackets and is a part of the `documentclass` command:

```
\documentclass[STYLE]{acmart}
```

Journals use one of three template styles. All but three ACM journals use the `acmsmall` template style:

- `acmsmall`: The default journal template style.
- `acmlarge`: Used by JOCCH and TAP.
- `acmtog`: Used by TOG.

The majority of conference proceedings documentation will use the `acmconf` template style.

- `acmconf`: The default proceedings template style.
- `sigchi`: Used for SIGCHI conference articles.
- `sigchi-a`: Used for SIGCHI "Extended Abstract" articles.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Table 2 from the paper

Classifier	Modality	Precision	Recall	F1
K-Nearest neighbor	metadata	0.607	0.654	0.63
Nearest mean classi?er	metadata	0.603	0.579	0.591
Decision tree	metadata	0.538	0.591	0.563
Logistic regression	metadata	0.548	0.609	0.578
SVM (Gaussian Kernel)	metadata	0.501	0.672	0.574
Bagging	metadata	0.604	0.662	0.631
Random Forest	metadata	0.559	0.593	0.576
AdaBoost	metadata	0.511	0.563	0.536
Gradient Boosting Tree	metadata	0.544	0.596	0.569
Naive Bayes	textual	0.545	0.987	0.702
K-Nearest neighbor	textual	0.549	0.844	0.666
SVM (Gaussian Kernel)	textual	0.547	1	0.707
K-Nearest neighbor	visual	0.582	0.636	0.608
Decision tree	visual	0.521	0.55	0.535
Logistic regression	visual	0.616	0.6	0.608
SVM (Gaussian Kernel)	visual	0.511	0.67	0.58
Random Forest	visual	0.614	0.664	0.638
AdaBoost	visual	0.601	0.717	0.654
Gradient Boosting Tree	visual	0.561	0.616	0.587
Logistic regression	audio	0.507	0.597	0.546
Gradient Boosting Tree	audio	0.56	0.617	0.58

- `sigplan`: Used for SIGPLAN conference articles.

2.1 Template Parameters

In addition to specifying the *template style* to be used in formatting your work, there are a number of *template parameters* which modify some part of the applied template style. A complete list of these parameters can be found in the *LaTeX User's Guide.*

Frequently-used parameters, or combinations of parameters, include:

- `anonymous,review`: Suitable for a "double-blind" conference submission. Anonymizes the work and includes line numbers. Use with the `\acmSubmissionID` command to print the submission's unique ID on each page of the work.
- `authorversion`: Produces a version of the work suitable for posting by the author.
- `screen`: Produces colored hyperlinks.

This document uses the following string as the first command in the source file:

```
\documentclass[sigconf]{acmart}
```

3 REPDORUCING TABLE 2

The second table we had to reproduce required two methods

3.1 scikit part 1

3.2 scikit part 2

Modifying the template — including but not limited to: adjusting margins, typeface sizes, line spacing, paragraph and list definitions, and the use of the `\vspace` command to manually adjust the vertical spacing between elements of your work — is not allowed.

Table 2 repdroduction

Classifier	Modality	Precision	Recall	F1	Best Feature
k-Nearest neighbor	audio	0.506	0.617	0.55	10
Decision tree	audio	0.62	0.657	0.622	4
SVM (Gaussian Kernel)	audio	0.51	0.737	0.543	13
Random Forest	audio	0.526	0.583	0.545	13
AdaBoost	audio	0.602	0.6	0.591	4
Gradient Boosting Tree	audio	0.595	0.657	0.619	6
k-Nearest neighbor	textual	0.523	0.77	0.619	1543
Decision tree	textual	0.547	0.823	0.653	116
Logistic regression	textual	0.533	0.617	0.558	3282
SVM (Gaussian Kernel)	textual	0.527	0.903	0.664	3282
Bagging	textual	0.615	0.753	0.658	3282
Random Forest	textual	0.559	0.82	0.654	3282
AdaBoost	textual	0.677	0.887	0.756	2375
Gradient Boosting Tree	textual	0.669	0.863	0.744	1007
Naive Bayes	textual	0.591	0.73	0.645	3211
k-Nearest neighbor	visual	0.614	0.677	0.641	757
Decision tree	visual	0.6	0.553	0.555	16
Logistic regression	visual	0.617	0.697	0.633	826
SVM (Gaussian Kernel)	visual	0.577	0.92	0.708	568
Bagging	visual	0.677	0.58	0.614	826
Random Forest	visual	0.561	0.607	0.563	198
AdaBoost	visual	0.606	0.6	0.592	826
Gradient Boosting Tree	visual	0.605	0.617	0.602	826
Naive Bayes	visual	0.548	0.687	0.588	826
k-Nearest neighbor	metadata	0.579	0.557	0.56	75
Decision tree	metadata	0.558	0.517	0.533	39
Logistic regression	metadata	0.569	0.54	0.536	22
SVM (Gaussian Kernel)	metadata	0.548	1.0	0.707	75
Random Forest	metadata	0.527	0.553	0.526	75
AdaBoost	metadata	0.664	0.54	0.576	75
Gradient Boosting Tree	metadata	0.541	0.567	0.548	75

Table 3 from the Paper

Stacking Strategy	Precision	Recall	F1
Voting (cv)	0.94	0.57	0.71
Label Stacking (cv)	0.72	0.86	0.78
Label Attribute Stacking (cv)	0.71	0.79	0.75
Voting (test)	0.62	0.8	0.7
Label Stacking (test)	0.62	0.9	0.73

Your document will be returned to you for revision if modifications are discovered.

4 REPRODUCING TABLE 3

The "`acmart`" document class requires the use of the "Libertine" typeface family. Your TeX installation should include this set of packages. Please do not substitute other typefaces. The "`lmodern`" and "`ltimes`" packages should not be used, as they will override the built-in typeface families.

5 TITLE INFORMATION

The title of your work should use capital letters appropriately - https://capitalizemytitle.com/ has useful rules for capitalization. Use the `title` command to define the title of your work. If your work has a subtitle, define it with the `subtitle` command. Do not insert line breaks in your title.

If your title is lengthy, you must define a short version to be used in the page headers, to prevent overlapping text. The `title` command has a “short title” parameter:

```
\title[short title]{full title}
```

6 AUTHORS AND AFFILIATIONS

Each author must be defined separately for accurate metadata identification. Multiple authors may share one affiliation. Authors’ names should not be abbreviated; use full first names wherever possible. Include authors’ e-mail addresses whenever possible.

Grouping authors’ names or e-mail addresses, or providing an “e-mail alias,” as shown below, is not acceptable:

```
\author{Brooke Aster, David Mehldau}
\email{dave,judy,steve@university.edu}
\email{firstname.lastname@phillips.org}
```

The `authornote` and `authornotemark` commands allow a note to apply to multiple authors — for example, if the first two authors of an article contributed equally to the work.

If your author list is lengthy, you must define a shortened version of the list of authors to be used in the page headers, to prevent overlapping text. The following command should be placed just after the last `\author{}` definition:

```
\renewcommand{\shortauthors}{McCartney, et al.}
```

Omitting this command will force the use of a concatenated list of all of the authors’ names, which may result in overlapping text in the page headers.

The article template’s documentation, available at https://www.acm.org/publications/proceedings-template, has a complete explanation of these commands and tips for their effective use.

Note that authors’ addresses are mandatory for journal articles.

7 RIGHTS INFORMATION

Authors of any work published by ACM will need to complete a rights form. Depending on the kind of work, and the rights management choice made by the author, this may be copyright transfer, permission, license, or an OA (open access) agreement.

Regardless of the rights management choice, the author will receive a copy of the completed rights form once it has been submitted. This form contains LaTeX commands that must be copied into the source document. When the document source is compiled, these commands and their parameters add formatted text to several areas of the final document:

- the “ACM Reference Format” text on the first page.
- the “rights management” text on the first page.
- the conference information in the page header(s).

Rights information is unique to the work; if you are preparing several works for an event, make sure to use the correct set of commands with each of the works.

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

The ACM Reference Format text is required for all articles over one page in length, and is optional for one-page articles (abstracts).

8 CCS CONCEPTS AND USER-DEFINED KEYWORDS

Two elements of the “acmart” document class provide powerful taxonomic tools for you to help readers find your work in an online search.

The ACM Computing Classification System — https://www.acm.org/publications/class-2012 — is a set of classifiers and concepts that describe the computing discipline. Authors can select entries from this classification system, via https://dl.acm.org/ccs/ccs.cfm, and generate the commands to be included in the LaTeX source.

User-defined keywords are a comma-separated list of words and phrases of the authors’ choosing, providing a more flexible way of describing the research being presented.

CCS concepts and user-defined keywords are required for for all articles over two pages in length, and are optional for one- and two-page articles (or abstracts).

9 SECTIONING COMMANDS

Your work should use standard LaTeX sectioning commands: `section`, `subsection`, `subsubsection`, and `paragraph`. They should be numbered; do not remove the numbering from the commands.

Simulating a sectioning command by setting the first word or words of a paragraph in boldface or italicized text is **not allowed.**

10 TABLES

The “`acmart`” document class includes the “`booktabs`” package — https://ctan.org/pkg/booktabs — for preparing high-quality tables.

Table captions are placed *above* the table.

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper “floating” placement of tables, use the environment **table** to enclose the table’s contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material are found in the *LaTeX User’s Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed output of this document.

To set a wider table, which takes up the whole width of the page’s live area, use the environment **table*** to enclose the table’s contents and the table caption. As with a single-column table, this wide table will “float” to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

11 MATH EQUATIONS

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

11.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin . . . \end` construction or with the short form `$ . . . $`. You can use any of the symbols and structures, from α to ω, available in LaTeX [?]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\rightarrow\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

11.2 Display Equations

A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\rightarrow\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_{0}^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

12 FIGURES

The "`figure`" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Your figures should contain a caption which describes the figure to the reader. Figure captions go below the figure. Your figures should **also** include a description suitable for screen readers, to assist the visually-challenged to better understand your work.

Figure captions are placed *below* the figure.

12.1 The "Teaser Figure"

A "teaser figure" is an image, or set of images in one figure, that are placed after all author and affiliation information, and before the body of the article, spanning the page. If you wish to have such a figure in your article, place the command immediately before the `\maketitle` command:

Figure 1: 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

```
\begin{teaserfigure}
  \includegraphics[width=\textwidth]{sampleteaser}
  \caption{figure caption}
  \Description{figure description}
\end{teaserfigure}
```

13 CITATIONS AND BIBLIOGRAPHIES

The use of BibTEX for the preparation and formatting of one's references is strongly recommended. Authors' names should be complete — use full first names ("Donald E. Knuth") not initials ("D. E. Knuth") — and the salient identifying features of a reference should be included: title, year, volume, number, pages, article DOI, etc.

The bibliography is included in your source document with these two commands, placed just before the `\end{document}` command:

```
\bibliographystyle{ACM-Reference-Format}
\bibliography{bibfile}
```

where "`bibfile`" is the name, without the "`.bib`" suffix, of the BibTEX file.

Citations and references are numbered by default. A small number of ACM publications have citations and references formatted in the "author year" style; for these exceptions, please include this command in the **preamble** (before "`\begin{document}`") of your LaTeX source:

```
\citestyle{acmauthoryear}
```

Some examples. A paginated journal article [?], an enumerated journal article [?], a reference to an entire issue [?], a monograph (whole book) [?], a monograph/whole book in a series (see 2a in spec. document) [?], a divisible-book such as an anthology or compilation [?] followed by the same example, however we only output the series if the volume number is given [?] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter

Table 2: Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

in a divisible book [?], a chapter in a divisible book in a series [?], a multi-volume work as book [?], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [?], a proceedings article with all possible elements [?], an example of an enumerated proceedings article [?], an informally published work [?], a doctoral dissertation [?], a master's thesis: [?], an online document / world wide web resource [? ? ?], a video game (Case 1) [?] and (Case 2) [?] and [?] and (Case 3) a patent [?], work accepted for publication [?], 'YYYYb'-test for prolific author [?] and [?]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [?]. Boris / Barbara Beeton: multi-volume works as books [?] and [?]. A couple of citations with DOIs: [? ?]. Online citations: [? ? ?]. Artifacts: [?] and [?].

14 ACKNOWLEDGMENTS

Identification of funding sources and other support, and thanks to individuals and groups that assisted in the research and the preparation of the work should be included in an acknowledgment section, which is placed just before the reference section in your document.

This section has a special environment:

```
\begin{acks}
...
\end{acks}
```

so that the information contained therein can be more easily collected during the article metadata extraction phase, and to ensure consistency in the spelling of the section heading.

Authors should not prepare this section as a numbered or unnumbered `\section`; please use the "`acks`" environment.

15 APPENDICES

If your work needs an appendix, add it before the "`\end{document}`" command at the conclusion of your source document.

Start the appendix with the "`appendix`" command:

```
\appendix
```

and note that in the appendix, sections are lettered, not numbered. This document has two appendices, demonstrating the section and subsection identification method.

16 SIGCHI EXTENDED ABSTRACTS

The "`sigchi-a`" template style (available only in LaTeX and not in Word) produces a landscape-orientation formatted article, with a wide left margin. Three environments are available for use with the "`sigchi-a`" template style, and produce formatted output in the margin:

- `sidebar`: Place formatted text in the margin.
- `marginfigure`: Place a figure in the margin.
- `margintable`: Place a table in the margin.

REFERENCES

[1] Cynthia C. S Liem and Bo Wang. *TUD-MMC at MediaEval 2016: Context of Experience task*, 2016.

[2] M. Riegler, M. Larson, C. Spampinato, P. Halvorsen, M. Lux, J. Markussen, K. Pogorelov, C. Griwodz, and H. Stensland. *Right inflight? A dataset for exploring the automatic prediction of movies suitable for a watching situation. In Proceedings of the 7th International Conference on Multimedia Systems*, pages 45:1–45:6. ACM, 2016.