1 Introduction

ACM's consolidated article template, introduced in 2017, provides a consistent LaTeX style for use across ACM publications, and incorporates accessibility and metadata-extraction functionality necessary for future Digital Library endeavors. Numerous ACM and SIG-specific LaTeX templates have been examined, and their unique features incorporated into this single new template.

If you are new to publishing with ACM, this document is a valuable guide to the process of preparing your work for publication. If you have published with ACM before, this document provides insight and instruction into more recent changes to the article template.

The "`acmart`" document class can be used to prepare articles for any ACM publication — conference or journal, and for any stage of publication, from review to final "camera-ready" copy, to the author's own version, with *very* few changes to the source.

2 Data Exploring

The "Trending YouTube Video Statistics" is selected as the dataset [**?**]. In general, the dataset focus on daily records of trending viedos on Youtube. It consists of informations of trending viedos in 10 countries (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India). Entries from each country are stored in a CSV file. For each entry, the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count are recorded, where video title, channel title, tags and description are text and others are digits. Among them, the title, tags, description, views, likes, dislikes and comment count are selected for further tasks.

2.1 Distributions of numerical data

The distributions of all numerical data in the dataset are obtained and those distributions of different countries shares similarity.

Table 1. Descriptive Statistics of Video Metrics (CA)

Statistic	Views	Likes	Dislikes	Comment Count
Count	39,585	39,585	39,585	39,585
Mean	1,169,234.01	40,596.94	2,058.69	5,159.72
Std. Dev.	3,437,842.10	134,596.73	19,312.58	21,899.59
Min	733	0	0	0
25%	149,715	2,395	104	442
Median	383,120	9,244	314	1,357
75%	983,139	29,670	976	3,821
Max	137,843,120	5,053,338	1,602,383	1,114,800

According to Table **??**, the basic statistics information of numerical data in Canada is provided. Both the mean and median values of views are much larger than those of likes, dislikes and comment count. Additionally, the standard deviations of all metrics are larger than their mean but smaller than ten times of the mean, which means their distributions are dispersive, such as the exponential distribution.

For each numerical metrics, their distributions are generated and the distributions in Canada are provided. And the probability distribution functions (PDF) obtained fromstimation (KDE) kernel density and exponential fitting are used. According to Figure **??**, Figure **??**, Figure **??**, Figure **??**, their distributions are exponential distributions, meaning that most of the data are small and close with few extremely high data. Therefore, all numerical data are preprocessed into log-scale in further tasks to ensure the models to converge.

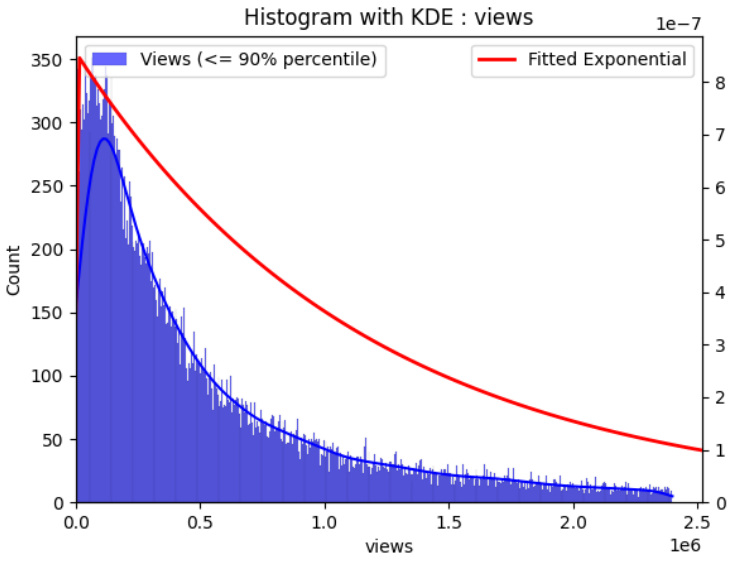

Figure 1. The distribution of views in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

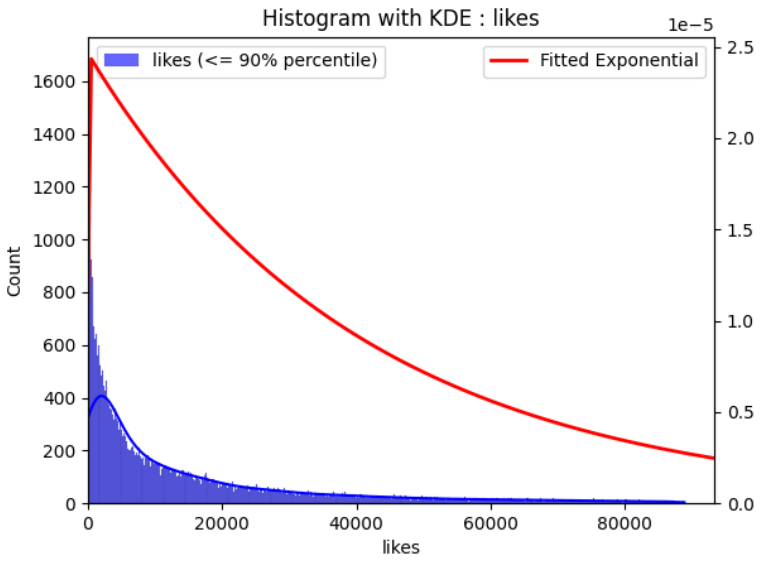

Figure 2. The distribution of likes in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

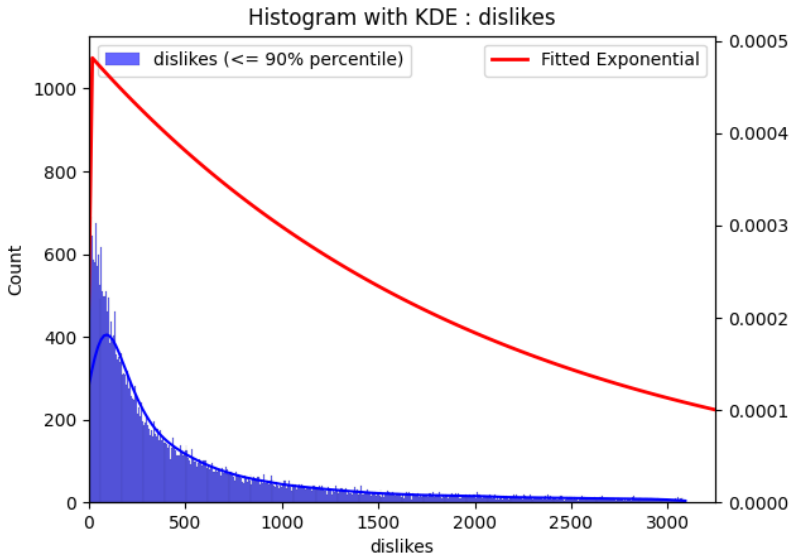

Figure 3. The distribution of dislikes in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

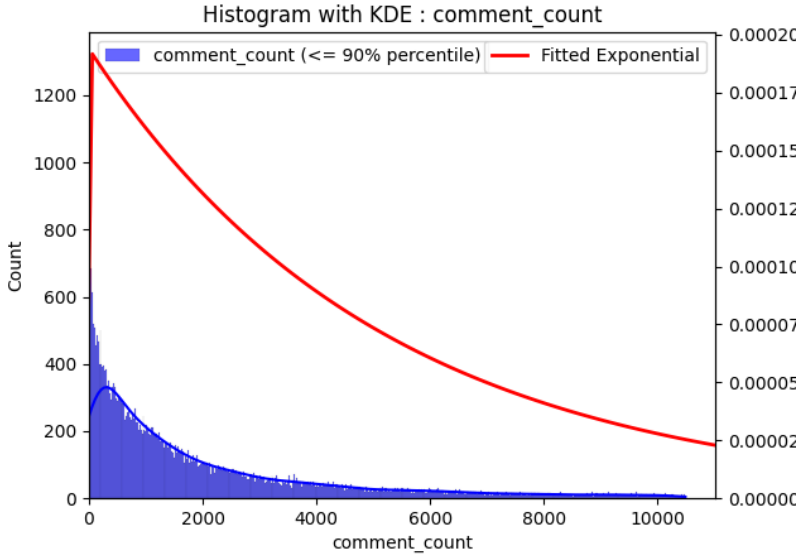

Figure 4. The distribution of comments count in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

2.2 Word Frequency

The frequency of words in the text data provided signification information about the video. After removing the irrelevant informations such as stopping words and urls, the word frequency of title, tags and descriptions are generated as the word clouds.

Figure 5. The word cloud of titles in Canada.

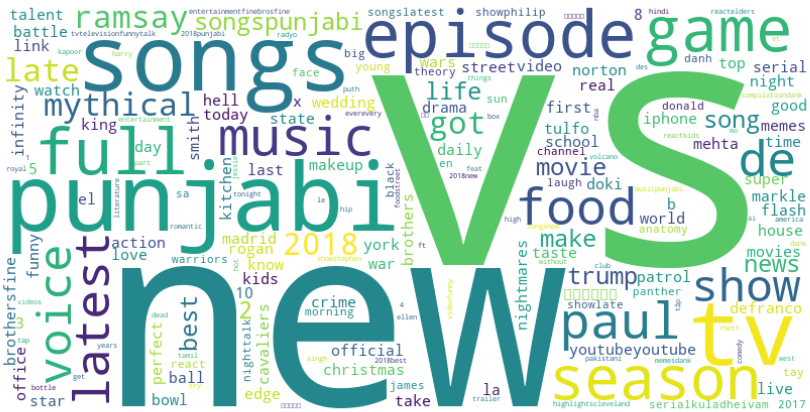

Figure 6. The word cloud of tags in Canada.

Figure 7. The word cloud of descriptions in Canada.

3 Task

Since all features selected are expected to share strong releations, each feature can be potentialy predicted by other features. Taking consideration of the real world application, the prediction of views of a viedo is selected as the task. Since views of a viedo can significantly indicate its popularity, the effective prediction can direct the creators while improve the recommendation methods of platforms.

As the prediction target is views, a regression model is expected. Initially, the simple logistic regression models with other numerical features such as likes, dislikes and comment count is selected as baselines. Further, the Term Frequency-Inverse Document Frequency (TF-IDF) approach is expected to perform well with text features, such as titles, tags and descriptions. The MSE, MAE and R^2 are selected for evaluation.

All numerical features are converted into log-scale and all text features are preprocessed by removing irrelevant informations.

4 Model

To predict the views of a viedo, all other metrics in the entry can be considered as features. Since the views, likes, dislikes and comment count share the same distributions, the regression models using likes, dislikes and comment count as inputs are expected to perfomr well. However, the prediction task can not be actually completed with these models since the views data is available if likes, dislikes and comment count are available. And using features to predict an existed data is meaningless.

The title, tags and descriptions of a viedo are available once it is uploaded. Therefore, these informations can be applied as features for prediction. Then the TF-IDF approach can be used, which provides the contribution of a word to text. The word with higher frequency in a sample text and lower frequency in the whole text is assigned with higher weight, and vice versa. The method focus on the releation between each signle word and the sample text. However, the context information is aborted and all words are considered as independent.

Introduced by [**?**], the self-attention based approach : Transformer, is proven to perform well in processing texts. It weights the importance of releations among different words, enabling it to extract the context and global information. In general, the model develops the encoder-decoder structure. In the encoder, inputs are converted into tokens, which are the minimum units of information. Then, linear layers are applied to map the input tokens into three matrixes: Query, Key and Value. And the attention weights are generated by the matrixes.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In the decoder, a prediction head is used to generate the prediction result from the weights.

To apply Transformer in the views prediction task, the features (title, tags and descriptions) are converted into tokens, which are the basic unit of the text. After that, a pre-trained Transformer model with signle head is adopted. In each epoch of training, the model's performance on the validation set is evaulated and early-stopping is adopted to avoid overfitting.

5 Literature

Similar prediction tasks are completed with the dataset, such as likes prediction, Category Prediciton [**? ? ?**]. Various approaches such as the random forest, XGBoost and linear regression are applied [**?**].

Although Transformer based approach archives the best performance, it is limited by huge time and space consumption since each token has interactions with all tokens. Although its self-attention mechanism can obtain the context information, the computation cost increases rapidly as the text seqence increasing. The Receptance-Weighted Key-Value (RWKV) model combines the RNN and self-attention, archives high performance on large scale tasks with acceptable costs.

6 Experiments and Results

Table 2. Performance Metrics of Models

Model	MSE	MAE	R^2
Single Feature (likes)	0.77	0.63	0.74
Single Feature (comments)	1.19	0.77	0.60
Single Feature (dislikes)	0.76	0.64	0.74
TF-IDF (tags)	0.43	0.33	0.86
TF-IDF-SVD (tags)	1.27	0.42	**0.58**
TF-IDF (description)	0.80	0.31	0.74
TF-IDF-SVD (description)	0.55	0.39	0.82
TF-IDF (title)	0.32	0.28	0.89
TF-IDF-SVD (title)	0.44	0.38	0.85
Transformer (tags)	0.44	0.38	0.85
Transformer (description)	0.30	0.35	0.90
Transformer (title)	**0.18**	**0.27**	0.93

Table **??** provides the performance of 12 different models. The dataset is likes, comments, dislikes, tags, description and titles in US, with size of 40739. After shuffle, 90% of the dataset is devided as the training set and 10% is the validation set. The metrics of MSE, MAE and R^2 are used.

The baselines are in three groups: regression models with signle features, TF-IDF models and TF-IDF-SVD models. For TF-IDF models, the max 20000 imprtant words are used. For TF-ID-SVD models, the Singular Value Decomposition approach is introduced for dimension reduction from 20000 to 5000.

As the result, the Transformer model with titles as features archives the best performance: 0.18 in MSE and 0.27 in MAE, proving the effectiveness of Transformer on texts prediction tasks. Furthermore, all models with titles perform better than other text features, indicating that the title of a video contributes more with its views than tags and descriptions.

7 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}
```

8 Authors and Affiliations

Each author must be defined separately for accurate metadata identification. As an exception, 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.

9 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.

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

10 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).

Table 3. 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

Table 4. Some Typical Commands

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

11 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.**

12 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 **??** 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 **??** 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.

Always use midrule to separate table header rows from data rows, and use it only for this purpose. This enables assistive technologies to recognise table headers and support their users in navigating tables more easily.

13 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.

13.1 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: $\lim_{n\to\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

13.2 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:

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

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.

14 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.

Figure 8. 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

Your figures should contain a caption which describes the figure to the reader.

Figure captions are placed *below* the figure.

Every figure should also have a figure description unless it is purely decorative. These descriptions convey what's in the image to someone who cannot see it. They are also used by search engine crawlers for indexing images, and when images cannot be loaded.

A figure description must be unformatted plain text less than 2000 characters long (including spaces). **Figure descriptions should not repeat the figure caption – their purpose is to capture important information that is not already provided in the caption or the main text of the paper.** For figures that convey important and complex new information, a short text description may not be adequate. More complex alternative descriptions can be placed in an appendix and referenced in a short figure description. For example, provide a data table capturing the information in a bar chart, or a structured list representing a graph. For additional information regarding how best to write figure descriptions and why doing this is so important, please see https://www.acm.org/publications/taps/describing-figures/.

14.1 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:

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

15 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 the command "`\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 in a divisible book [**?**], a chapter in a divisible book in a series [**?**], a multi-volume work as book [**?**], a couple of articles 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 couple of preprints [**? ?**], 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 [**?**].

16 Acknowledgments

Identification of funding sources and other support, and thanks toAbril07 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.

17 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.

18 Multi-language papers

Papers may be written in languages other than English or include titles, subtitles, keywords and abstracts in different languages (as a rule, a paper in a language other than English should include an English title and an English abstract). Use `language=...` for every language used in the paper. The last language indicated is the main language of the paper. For example, a French paper with additional titles and abstracts in English and German may start with the following command

```
\documentclass[sigconf, language=english, language=german,
               language=french]{acmart}
```

The title, subtitle, keywords and abstract will be typeset in the main language of the paper. The commands `\translatedXXX`, `XXX` begin title, subtitle and keywords, can be used to set these elements in the other languages. The environment `translatedabstract` is used to set the translation of the abstract. These commands and environment have a mandatory first argument: the language of the second argument. See `sample-sigconf-i13n.tex` file for examples of their usage.

19 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.

Acknowledgments

To Robert, for the bagels and explaining CMYK and color spaces.

References

[] [n. d.]. *Trending YouTube Video Statistics*. https://www.kaggle.com/datasets/datasnaek/youtube-new/data
[] [n. d.]. *youtube likes prediction*. https://www.kaggle.com/code/hetulmehta/youtube-likes-prediction
[] [n. d.]. *YouTube Likes Prediction using 3 ML models*. https://www.kaggle.com/code/sanketsonu/youtube-likes-prediction-using-3-ml-models
[] [n. d.]. *Youtube Videos Category Prediciton*. https://www.kaggle.com/code/emirfarukerman/98-accuracy-youtube-videos-category-prediciton
[] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

A Research Methods

A.1 Part One

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna, id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.

A.2 Part Two

Etiam commodo feugiat nisl pulvinar pellentesque. Etiam auctor sodales ligula, non varius nibh pulvinar semper. Suspendisse nec lectus non ipsum convallis congue hendrerit vitae sapien. Donec at laoreet eros. Vivamus non purus placerat, scelerisque diam eu, cursus ante. Etiam aliquam tortor auctor efficitur mattis.

B Online Resources

Nam id fermentum dui. Suspendisse sagittis tortor a nulla mollis, in pulvinar ex pretium. Sed interdum orci quis metus euismod, et sagittis enim maximus. Vestibulum gravida massa ut felis suscipit congue. Quisque mattis elit a risus ultrices commodo venenatis eget dui. Etiam sagittis eleifend elementum.

Nam interdum magna at lectus dignissim, ac dignissim lorem rhoncus. Maecenas eu arcu ac neque placerat aliquam. Nunc pulvinar massa et mattis lacinia.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.