NLP & Machine Learning Applied: Video Game Reviews

Matthew D. Mulholland
Montclair State University
Montclair, NJ
mulhollandm2@montclair.edu

Abstract

Despite the ambiguity of the concept, much research has been done in the area of detecting "fake" reviews. It is, however, often difficult to build corpora containing reviews that are definitively "fake" or "true". It is simpler to reframe the issue in terms of the amount of experience a reviewer has with a product: given a review, can we tell anything about the level of experience the reviewer has with the reviewed product? In this exploratory paper, I will detail a research project whose aim was to relate video game reviews to proxies for reviewer experience, such as number of hours played, number of times marked as helpful, and other related review/user attributes, using natural language processing and machine learning techniques. The ultimate end is to produce a capability for ranking or filtering reviews, which could be used in addition to or in place of other fake review or spam filtering algorithms. A less grand aim of the project was to scrape review data from the Steam video game website and make it publicly available. This data will be described at length.

1 Credits

I would like to thank both Janette Martinez and Emily Olshefski for their help in the initial iteration of this work as part of a class project. They helped lay out the problem, make decisions regarding the source of the data and the games for which data was collected, and also write some of the preprocessing code. Some sections of this paper build on the final paper for that class project, which they took a lead in writing.

2 Introduction

Ground truth in the realm of deception detection is information that can be verified or denied. One area where deception detection is applied is in the detection of fake reviews. In this case, reviews that are known to be fake (by some external means) are compared to reviews that are believed to have been written in good faith. However, defining reviews from either category is a difficult matter and corpora of fake/real reviews are often constrained in that, for any given review, it could be impossible to determine the category. In this paper, I propose a fundamental reframing of the issue: the problem should not be about whether or not a given review is fake, but rather it should be about determining the amount of experience a reviewer has with the product being reviewed. Reviews for which the reviewer has little to no experience with the product – whether they have been written in bad faith or simply from a relatively uninformed perspective – could be distinguished from reviews for which the reviewer does have experience with the product being reviewed. Further, different levels of experience – say, moderate or high – could be distinguished from one another. A system that could accurately predict the amount of experience a reviewer has given only the review text or some combination of the review text and some other attributes of the review/reviewer, such as the number of times a review has been marked as helpful, could be useful in a production environment as a review filtering algorithm and/or a review sorting algorithm.

A way of determining a reviewer's experience with a product that could be used to train a system is a prerequisite of this line of research. In the realm of video game reviews, this is a feasible prerequisite: the Steam online video game platform, for instance, keeps a record of the number of hours each user has played each game and, thus, when a user submits a review of a video game, this information is presented alongside the review. It is true that the meaning of this measure is not completely straightforward: a user could have played a particular game for many hours outside the Steam platform and these hours would not be included in the record of that user's playing time in the online platform. However, I believe that it is a reasonable assumption to make that the majority of the values recorded by the platform will be accurate and/or that the degree to which a player has played a particular game outside the confines of the online platform will be similar across players submitting reviews for the same game. Thus, the relative amount of experience embodied by the number of hours played could still be a valid piece of information. The window into a user's experience with a product that is afforded by the amount of time a user has used a product, however, is somewhat unique to the case of video games: for other products, it might be impossible or pointless to attempt to record the amount of time the user used the product. For example, there is no way to keep a record of the amount of time a user has spent using a vacuum short of conducting an experiment and recording such information in person. However, if experience could be successfully modelled in the case of video games, perhaps the models could be generalized to cover whole categories of products and, thus, the situation here would apply to much more than video games. Furthermore, there are other indications of a user's experience with a product, such as the number of times a review has been marked as helpful.

In this paper, I look to the hours played by the reviewer as the ground truth. The data we are basing the ground truth upon are reviews of games published on the online video game platform Steam. The ground truth in our case is the hours played values, which are provided by Steam for all reviews and which are derived from the players' own online game playing statistics (i.e., not on user-reported values). The information gathered by Steam is a good example of ”external” ground truth, along with the fact that this is real-world data. The idea is that the higher the hours played values are, the more trustworthy the reviews are. Naturally, if a reviewer has played more hours of a game, then he/she has a better understanding of the game and, in turn, can write a more nuanced and accurate review of a game or a review that exhibits information that can only be acquired through usage.

Data was collected from the video game platform Steam, available or PC, Mac, and Linux, due its popularity as a gaming platofrm as well as availability of data. We developed a web-scraping method in order to build a corpus of reviews. Reviews from the top 11 most popular games were scraped from the Steam website. Number of hours played were also collected in conjunction with the review text associated with them. After pruning the data – filtering out non-English reviews, reviews that had close to no content, etc., training/test set partitions were built. Using Weka and a basic bag-of-words approach in order to get some initial results, we built regression models with the SMOreg machine learning algorithm. This work is the first steps in using novel and more comprehensive data not otherwise used in similar studies.

3 General Instructions

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, authors' names and complete addresses, which must be centered at the top of the first page, and any full-width figures or tables (see the guidelines in Subsection 3.5). **Type single-spaced.** Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section 6. Do not number the pages.

3.1 Electronically-available resources

We strongly prefer that you prepare your PDF files using LaTeX with the official ACL 2015 style file (acl2015.sty) and bibliography style (acl.bst). These files are available at `http://acl2015.org`. You will also find the document you are currently reading (acl2015.pdf) and its LaTeX source code (acl2015.tex) on this website.

You can alternatively use Microsoft Word to produce your PDF file. In this case, we strongly recommend the use of the Word template file (acl2015.dot) on the ACL 2015 website (`http://acl2015.org`). If you have an option, we recommend that you use the LaTeX2e version. If you will be using the Microsoft Word template, we suggest that you anonymize your source file so that the pdf produced does not retain your identity. This can be done by removing any personal information from your source document properties.

3.2 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). PDF files are usually produced from LaTeX using the *pdflatex* command. If your version of LaTeX produces Postscript files, you can convert these into PDF using *ps2pdf* or *dvipdf*. On Windows, you can also use Adobe Distiller to generate PDF.

Please make sure that your PDF file includes all the necessary fonts (especially tree diagrams, symbols, and fonts with Asian characters). When you print or create the PDF file, there is usually an option in your printer setup to include none, all or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. **Before sending it, test your PDF by printing it from a computer different from the one where it was created.** Moreover, some word processors may generate very large PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying “Output to a file”, then convert the file to PDF.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the paper. When working with `dvips`, for instance, one should specify `-t a4`. Or using the command `\special{papersize=210mm,297mm}` in the latex preamble (directly below the `\usepackage` commands). Then using `dvipdf` and/or `pdflatex` which would make it easier for some.

Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs as soon as possible.

3.3 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper are:

- Left and right margins: 2.5 cm
- Top margin: 2.5 cm
- Bottom margin: 2.5 cm
- Column width: 7.7 cm
- Column height: 24.7 cm
- Gap between columns: 0.6 cm

Papers should not be submitted on any other paper size. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

3.4 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

Type of Text	**Font Size**	**Style**
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

3.5 The First Page

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. Use the two-column format only when you begin the abstract.

Title: Place the title centered at the top of the first page, in a 15-point bold font. (For a complete guide to font sizes and styles, see Table 1) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author's names(s), and the affiliation on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (e.g., use "Schlangen" not "SCHLANGEN"). Do not format title and section headings in all capitals as well except for proper names (such as "BLEU") that are conventionally in all capitals. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronical paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6 cm on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The abstract text should be in 10 point font.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers.

Indent when starting a new paragraph. Use 11 points for text and subsection headings, 12 points for section headings and 15 points for the title.

3.6 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsubsections.

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Append lowercase letters to the year in cases of ambiguity. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972). Also refrain from using full citations as sentence constituents. We suggest that instead of

> "(Gusfield, 1997) showed that ..."

you use

> "Gusfield (1997) showed that ..."

If you are using the provided LaTeX and BibTeX style files, you can use the command `\newcite` to get "author (year)" citations.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g.,

> "We previously showed (Gusfield, 1997) ..."

should be avoided. Instead, use citations such as

> "Gusfield (1997) previously showed ... "

Please do not use anonymous citations and do not include acknowledgements when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (Association for Computing Machinery, 1983).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

3.7 Footnotes

Footnotes: Put footnotes at the bottom of the page and use 9 points text. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2]

3.8 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions of the figures and tables below the body, using 11 point text.

4 XML conversion and supported LaTeX packages

Following ACL 2014 we will also we will attempt to automatically convert your LaTeX source files to publish papers in machine-readable XML with semantic markup in the ACL Anthology, in addition to the traditional PDF format. This will allow us to create, over the next few years, a growing corpus of scientific text for our own future research, and picks up on recent initiatives on converting ACL papers from earlier years to XML.

We encourage you to submit a ZIP file of your LaTeX sources along with the camera-ready version of your paper. We will then convert them to XML automatically, using the LaTeXML tool (`http://dlmf.nist.gov/LaTeXML`). LaTeXML has *bindings* for a number of LaTeX packages, including the ACL 2015 stylefile. These bindings allow LaTeXML to render the commands from these packages correctly in XML. For best results, we encourage you to use the packages that are officially supported by LaTeXML, listed at `http://dlmf.nist.gov/LaTeXML/manual/included.bindings`

5 Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration "translation".

6 Length of Submission

Long papers may consist of up to 8 pages of content, plus two extra pages for references. Short papers may consist of up to 4 pages of content, plus two extra pages for references. Papers that do not conform to the specified length and formatting requirements may be rejected without review.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

[1] This is how a footnote should appear.

[2] Note the line separating the footnotes from the text.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.