Corpus-based Deception Detection in Online Mafia using Neural Models

University College London, March 2020

Tom Grigg tom.grigg.19@ucl.ac.uk

Kelvin Ng kelvin.ng.19@ucl.ac.uk

Ritwick Sundar ritwick.sundar.19@ucl.ac.uk

Nicolas Ward nicolas.ward.19@ucl.ac.uk

Abstract

The aim of this project is to investigate various neural models to predict the underlying roles of players in online Mafia, a deceptive conversation game (also referred to as "Werewolf"). Our research should lead to the implementation of a deception detection algorithm capable of assessing the likelihood of each player being part of either the Mafia or the Villagers. The results of our paper should also offer a quantitative analysis of current findings in the field. Our work will use the Mafiascum dataset[1], a collection of over 700 games of Mafia in which players are randomly assigned either deceptive or non-deceptive roles and interact via forum postings.

1 Introduction

1.1 Scope

Automatic deception detection is a challenging problem in written human interactions, which has tremendous legal, political and social implications. Difficulties in text-based deception detection include the fact that human performance on the task is often at chance, that the signal in available data is dim, and that deception is a complex human behaviour whose manifestations depend on non-textual context, intrinsic properties of the deceiver (such as education, linguistic competence, and the nature of the intention), and specifics of the deceptive act (e.g., lying vs. fabricating).

We believe that deceivers and truth tellers display different linguistic patterns and singularities, and that those nuances can be intercepted by deep learning architectures. In this project, we attempt to design a system that is sensitive enough to discern the intricacies of deceptive language, yet robust enough to be replicated to other contexts. Due to the limited amount of deception corpuses available and the elusive nature of the problem, we approach our challenge in the context of online Mafia, a deceptive conversation game also referred to as Werewolf. We investigate various neural network architectures to predict the underlying role of each player in a given game. We show that our hybrid approach can improve a corpus-based deep learning model.

The findings of this project could contribute to the development of future research in emerging areas of linguistics and machine learning, with many promising applications ranging from measuring the accuracy of online reviews and combatting fake news to assessing the reliability of statements made in criminal trials based on tribunal hearing transcripts.

1.2 The Mafiascum Dataset

Although deception takes place in a large number of datasets, such as court transcripts (Perez-Rosas et al., 2015), the Enron email corpus (Keila and Skillicorn, 2005), and laboratory experiments (Perez-Rosas and Mihalcea, 2015), most public labelled datasets to date focus on single-sentence non-interactive deception. In this work, we use a dataset compiled by Ruiter and Kachergis (add reference) containing over 700 games of Mafia played on the Internet forum Mafiascum. The data consists of over 9000 documents, each containing all messages written by a single player in a single game. The average document contains 3940 words.

1.3 The Game of Mafia

Mafia (or Werewolf) is a social deduction game which models a conflict between two groups: an informed minority (the mafia), and an uninformed majority (the villagers). At the start of the game, each player is secretly assigned a role affiliated with one of these teams. The game has two alternating phases: during the night, the mafia covertly kills other players; during the day, surviving play-

[1] Link to the Mafiascum dataset: https://bitbucket.org/bopjesvla/thesis/src/master/

ers debate the identities of players and vote to eliminate a suspect. The game continues until one group achieves its win condition; for the villagers, this means eliminating all mafia players, while for the mafia this means reaching numerical parity with the villagers.

Mafia has been adopted by a large number of online communities. On the Internet, Days are usually played out in a public forum thread or chat room and the Mafia decide on the eliminated player by private messages.

In addition to its entertaining dynamics, the main reason for our interest in Werewolf is that, by design, mafia players have a strong incentive to conceal their identity through deception and false claims. In the following, we attempt to design a system capable of picking up on the cues and singularities involved in these deceptive scenarios.

2 Related work

To date, several studies have explored the identification of deceptive content in a variety of domains, including online dating, forums, social networks and consumer reviews. Until the 1970s, research aimed at detecting deception was largely focused on finding nonverbal cues (such as body language and facial expressions) for use in face-to-face interactions. In 1971, after Mehrabian found that slow and sparse speech can indicate deception, systemic research into linguistic cues to deception took off.

In 2003, Burgoon et al. found that the multivariate analysis of indicators of complexity at both the sentence level (simple sentences, long sentences, short sentences, sentence complexity, number of conjunctions, average-words-per-sentence) and vocabulary level (vocabulary complexity, number of big words, average-syllables-per-word) did not produce overall multivariate effects, but several individual variables did show the effects of deception condition. They showed that deceivers had significantly fewer long sentences, average-words-per-sentence and sentence complexity than truth tellers. This meant their language was less complex and easier to comprehend.

In a meta-analysis of 79 linguistic cues to deception from 44 studies, Hauch et al. (2015) found that relative to truth-tellers, liars experienced greater cognitive load, expressed more negative emotions, distanced themselves more from events by expressing fewer sensory-perceptual words, and referred less often to cognitive processes.

These works have shown the effectiveness of features derived from text analysis, which frequently includes basic linguistic representations such as n-grams and sentence counts statistics (Mihalcea and Strapparava, 2009; Ott et al., 2011) and also more complex linguistic features derived from syntactic context-free grammar trees and part-of-speech tags (Feng et al., 2012; Xu and Zhao, 2012).

One of the closest and most recent explorations of deception detection we are aware of (and our main inspiration for this work) was conducted by Ruiter and Kachergis (2018), who scraped Mafiascum's Normal Game archives to create a publicly available dataset of player conversations. They hand-picked linguistic features based on prior deception research and a set of average word vectors enriched with subword information that correlated with truthful or deceptive roles, such as word count per 24 hours, ratio of first/second/third-person pronouns, and ratio of negative emotion words. A logistic regression classifier fit on a combination of the feature sets achieved an AUROC of 0.68 on 5000+ word documents. According to their findings, of the six features said to be positively correlated with deception in their primary meta-analysis, only two (sentence length and ratio of third-person pronouns) had the same direction. This suggests that the strength of linguistic cues is highly context-dependent. The authors also note that linear models trained on average word vectors are unlikely to generalize across contexts.

Throughout this work, our aim is to build upon Ruiter and Kachergis's approach by using neural architectures, which have the considerable advantage of non-linearity. We expect these models to intercept more nuanced cues by accounting for context, co-references and variable semantic relationships, among other things. In the following, we formalize our approach.

3 Methods

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 3.6). **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 5. Pages are numbered for initial submission. However, **do not number the pages in the camera-ready version**.

By uncommenting `\aclfinalcopy` at the top of this document, it will compile to produce an example of the camera-ready formatting; by leaving it commented out, the document will be anonymized for initial submission. When you first create your submission on softconf, please fill in your submitted paper ID where `***` appears in the `\def\aclpaperid{***}` definition at the top.

The review process is double-blind, so do not include any author information (names, addresses) when submitting a paper for review. However, you should maintain space for names and addresses so that they will fit in the final (accepted) version. The ACL 2019 LaTeX style will create a titlebox space of 2.5in for you when `\aclfinalcopy` is commented out.

3.1 The Ruler

The ACL 2019 style defines a printed ruler which should be presented in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document without the provided style files, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\aclfinalcopy` command in the document preamble.)

Reviewers: note that the ruler measurements do not align well with lines in the paper – this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. In most cases one would expect that the approximate location will be adequate, although you can also use fractional references (*e.g.*, the first paragraph on this page ends at mark 108.5).

3.2 Electronically-available resources

ACL provides this description in LaTeX2e (`acl2019.tex`) and PDF format (`acl2019.pdf`), along with the LaTeX2e style file used to format it (`acl2019.sty`) and an ACL bibliography style (`acl_natbib.bst`) and example bibliography (`acl2019.bib`). These files are all available at `http://acl2019.org/downloads/acl2019-latex.zip`. We strongly recommend the use of these style files, which have been appropriately tailored for the ACL 2019 proceedings.

3.3 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.

3.4 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact

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
subsection titles	11 pt	bold
document text	11 pt	
captions	10 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

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.

3.5 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.

3.6 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 “Mitchell” not “MITCHELL”). 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: Indent when starting a new paragraph, about 0.4 cm. Use 11 points for text and subsection headings, 12 points for section headings and 15 points for the title.

3.7 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 sub-

Command	Output
{\"a}	ä
{\^e}	ê
{\`i}	ì
{\.I}	İ
{\o}	ø
{\'u}	ú
{\aa}	å
{\c c}	ç
{\u g}	ğ
{\l}	ł
{\~n}	ñ
{\H o}	ő
{\v r}	ř
{\ss}	ß

Table 2: Example commands for accented characters, to be used in, *e.g.*, BIBTEX names.

sections 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 (**?**) or, if the author's name appears in the text itself, as Gusfield (**?**). Using the provided LATEX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (**??**); this is accomplished with the provided style using commas within the `\cite` command, *e.g.*, `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (**?**), but write as in (**?**) when more than two authors are involved. Collapse multiple citations as in (**??**). Also refrain from using full citations as sentence constituents.

We suggest that instead of

> "(**?**) showed that ..."

you use

> "Gusfield (**?**) showed that ..."

If you are using the provided LATEX and BibTEX style files, you can use the command `\citet` (cite in text) to get "author (year)" citations.

You can use the command `\citealp` (alternative cite without parentheses) to get "author year" citations (which is useful for using citations within parentheses, as in **?**).

If the BibTEX file contains DOI fields, the paper title in the references section will appear as a hyperlink to the DOI, using the hyperref LATEX package. To disable the hyperref package, load the style file with the `nohyperref` option:
`\usepackage[nohyperref]{acl2019}`

Compilation Issues: Some of you might encounter the following error during compilation:

"`\pdfendlink` *ended up in different nesting level than* `\pdfstartlink`."

This happens when `pdflatex` is used and a citation splits across a page boundary. To fix this, the style file contains a patch consisting of the following two lines: (1) `\RequirePackage{etoolbox}` (line 454 in `acl2019.sty`), and (2) A long line below (line 455 in `acl2019.sty`).

If you still encounter compilation issues even with the patch enabled, disable the patch by commenting the two lines, and then disable the `hyperref` package (see above), recompile and see the problematic citation. Next rewrite that sentence containing the citation. (See, *e.g.*, `http://tug.org/errors.html`)

Digital Object Identifiers: As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. As of 2017, we are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus, please ensure that you use BibTEX records that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at `http://aclanthology.info/`.

As examples, we cite (**?**) to show you how papers with a DOI will appear in the bibliography. We cite (**?**) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

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 (**?**) ..."

should be avoided. Instead, use citations such as

> "**?** (**?**) previously showed ... "

Any preliminary non-archival versions of submitted papers should be listed in the submission form but not in the review version of the paper. ACL 2019 reviewers are generally aware that authors may present preliminary versions of their work in other venues, but will not be provided the list of previous presentations from the submission form.

output	natbib	previous ACL style files
(?)	`\citep`	`\cite`
?	`\citet`	`\newcite`
(?)	`\citeyearpar`	`\shortcite`

Table 3: Citation commands supported by the style file. The citation style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

Please do not use anonymous citations and do not include 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. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. By using a .bib file, as in this template, this will be automatically handled for you. See the `\bibliography` commands near the end for more.

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* (**?**). 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* (**?**).

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.

- Example citing an arxiv paper: (**?**).
- Example article in journal citation: (**?**).
- Example article in proceedings, with location: (**?**).
- Example article in proceedings, without location: (**?**).

See corresponding .bib file for further details.

Submissions should accurately reference prior and related work, including code and data. If a piece of prior work appeared in multiple venues, the version that appeared in a refereed, archival venue should be referenced. If multiple versions of a piece of prior work exist, the one used by the authors should be referenced. Authors should not rely on automated citation indices to provide accurate references for prior and related work.

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**.

3.8 Footnotes

Footnotes: Put footnotes at the bottom of the page and use 9 point font. They may be numbered or referred to by asterisks or other symbols.[2] Footnotes should be separated from the text by a line.[3]

3.9 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 10 point text. Captions should be placed below illustrations. Captions that are one line are centered (see Table 1). Captions longer than one line are left-aligned (see Table 2). Do not overwrite the default caption sizes. The acl2019.sty file is compatible with the caption and subcaption packages; do not add optional arguments.

3.10 Accessibility

In an effort to accommodate people who are color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. A simple criterion: All curves and points in your figures should be clearly distinguishable without color.

[2] This is how a footnote should appear.

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

4 Experiments

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".

5 Results and Discussion

The ACL 2019 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages of content plus unlimited pages for references – so that reviewers' comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references. For both long and short papers, all illustrations and tables that are part of the main text must be accommodated within these page limits, observing the formatting instructions given in the present document. Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

ACL 2019 does encourage the submission of additional material that is relevant to the reviewers but not an integral part of the paper. There are two such types of material: appendices, which can be read, and non-readable supplementary materials, often data or code. Do not include this additional material in the same document as your main paper. Additional material must be submitted as one or more separate files, and must adhere to the same anonymity guidelines as the main paper. The paper must be self-contained: it is optional for reviewers to look at the supplementary material. Papers should not refer, for further detail, to documents, code or data resources that are not available to the reviewers. Refer to Appendix A and Appendix B for further information.

Workshop chairs may have different rules for allowed length and whether supplemental material is welcome. As always, the respective call for papers is the authoritative source.

Conclusion

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.

Preparing References:
Include your own bib file like this:
`\bibliographystyle{acl_natbib}`
`\bibliography{acl2019}`

where `acl2019` corresponds to a acl2019.bib file.

A Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here. Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.