Leveraging Speaker Context for Natural Language Processing

Samee Ibraheem

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-242
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-242.html

December 1, 2023

Acknowledgement

This thesis is adapted from Investigating the Behavior of Malicious Actors Through the Game of Mafia, Putting the Con in Context: Identifying Deceptive Actors in the Game of Mafia, and a paper currently in submission at the time of submitting this dissertation. My co-authors, Vael Gates, John DeNero, Tom Griffiths, Gaoyue Zhou, and Risham Sidhu have all given me permission to include these works as part of my dissertation.

Leveraging Speaker Context for Natural Language Processing

by

Samee Omotayo Ibraheem

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John DeNero, Chair
Professor Marti Hearst
Professor David Bamman

Summer 2022

Leveraging Speaker Context for Natural Language Processing

Abstract

Leveraging Speaker Context for Natural Language Processing

by

Samee Omotayo Ibraheem

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor John DeNero, Chair

Neural networks have allowed for a host of advances in Natural Language Processing (NLP), from text classification to machine translation. These applications have demonstrated the ability to capture the effects of properties such as sentiment and politeness on language usage through computational means. However, using NLP to examine the effects of contextual information in relation to the intrinsic features of one's identity or the extrinsic features of one's conversational role is still an active area of research. This thesis focuses on modeling the effects of speaker attributes on language, looking at applications that are designed to help improve the safety of users in the digital world. Gender is a personal characteristic that people might not wish to share online but that can be determined by one's language use. We first examine how intrinsic speaker attributes affect language by attempting to obfuscate the gender of users on Reddit. Detecting deceptive actors in online interactions is also important for user security. We next explore the effect of extrinsic speaker attributes on language through the game of Mafia, in which participants may take on either an honest or a deceptive role. Through these analyses, we demonstrate that there are linguistic differences based on a person's role or identity, indicating that these aspects of an entity might be identified through their linguistic behavior. In addition to providing insight on how such entities use language in accordance with these features, these applications have implications for real-life communication paradigms, providing possible avenues for hiding aspects of one's identity or discovering aspects of another's.

To my treasured friends and chosen family

Contents

List of Figures

List of Tables

Acknowledgments

There are so many people that I would like to thank for their support throughout my PhD.

First of all, I am truly grateful to my advisor, Professor John DeNero, for his steady guidance in research and teaching, his incredible openness to my unusual background and ideas, and his kind encouragement in times of both hardship and triumph. Thanks to your example as an inspiring researcher, a caring teacher, and an amazing human being, I have learned and grown so much in the past six years.

I would also like to thank the other members of my dissertation committee, Professor Marti Hearst and Professor David Bamman, for their invaluable feedback and insights that have helped shape this research, as well as Professor Tom Griffiths, who similarly provided wonderful advice and direction as an additional member of my qualifying committee.

This dissertation would not have been possible without the help of my awesome research collaborators. Thank you to Professor John DeNero, Professor Tom Griffiths, Gaoyue Zhou, Risham Sidhu, Vael Gates, Aida Nematzadeh, and the Dallinger team for your enjoyable cooperation, your indispensable contribution, and your patient communication on these projects, as well as to Dan Liebling, Nick Altieri, Sherdil Niyaz, Ian McAulay, Saam Zahedian, Kazu Kogachi, and Huihan Liu for your fun and fulfilling collaborations earlier in my research career.

I have also benefited greatly from being a part of the Berkeley NLP community. I am especially thankful for the wonderful conversations and lunches that I have been able to have with Professor Dan Klein, David Gaddy, Katie Stasaski, Philippe Laban, Daniel Fried, Nick Altieri, Mitchell Stern, Nikita Kitaev, Max Rabinovich, Katia Patkin, Lucy Li, Cathy Chen, Rudy Corona, Nick Tomlin, Kevin Lin, Eric Wallace, Jessy Lin, Ruiqi Zhong, Steven Cao, and Kevin Yang.

There are also those in the wider Berkeley Artificial Intelligence Research (BAIR), Electrical Engineering and Computer Science (EECS), and College of Engineering (CoE) communities who have provided me assistance throughout my PhD, specifically Angie Abbatecola, Roxana Infante, Ami Katagiri, Shirley Salanio, Audrey Sillers, Jean Nguyen, Angela Waxman, and Meltem Erol.

I also wish to acknowledge the internship mentors who I was fortunate to be able to work with during my PhD. Thank you to Dan Liebling, Naveen Arivazhagan, Markus Freitag, and Mingxiang Zhu for providing unique perspectives and allowing me to explore different avenues for research.

I am further indebted to the various resources that I was afforded for my PhD, specifically through the Berkeley EECS Excellence Award, the Berkeley Chancellor's Fellowship for Graduate Study, the National Science Foundation Graduate Research Fellowship, and an NVIDIA Corporation GPU grant.

Before coming to Berkeley, I was able to train under marvelous mentors as an undergraduate at Harvard, including Professor Randy Buckner, Professor Emery Brown, Doctor Seun

Akeju, Professor David Malan, Professor Jelani Nelson, and Professor Haim Sompolinsky, who all helped foster my progression in research and teaching.

In addition to the many people who contributed to my development as a researcher, I have also had the pleasure of making formative connections through a variety of organizations during my time at Berkeley. I was blessed to be a part of several mentorship programs, including Graduate Pathways to STEM (GPS), the Graduate Minority Outreach, Recruitment, and Retention (GMORR) Mentorship Project, Girls Who Code, Getting into Graduate School (GiGS), the BAIR Undergraduate Mentoring Program, and the African-American Student Development Mentorship Program. I am glad that I got the opportunity to mentor students at all different stages of their education, and I hope that they gained as much as I have from our interactions. I am also especially appreciative of the opportunity to be a part of Black Graduate Engineering and Science Students (BGESS), Queers in Computer Science and Engineering (QICSE), T-Cal, and Nikkei Choral Ensemble (NiCE).

Finally, I would like to give a huge thank you to all of my friends, but especially to Sharon Zhou, Nisreen Shiban, Tanya Lee, Ben Mehlow, Jaimie Swartz, Emma Jiang, Maria Palaroan, Jessica Qian, Victoria Cheng, Care He, Pelagie Elimbi, and Thurston Dang, as well as to my family, Shakirat (Mom), Sarafa (Dad), Shafeeq, Shareef, and Saleem Ibraheem. I really appreciate you all and am lucky to have shared this great journey together.

Chapter 1

Introduction

Language is a powerful tool that allows not only for the communication of content between interlocutors, but also for the communication of a speaker's context. For example, if a person says "I am a happy woman", she not only states something about her current emotional state, but also something about one of her personal attributes, specifically her gender. In addition to such intrinsic speaker attributes, which are the properties of a person that are generally fixed over time, extrinsic speaker attributes that are specific to one's situational context may also be conveyed in a similar manner. For instance, if one were to state "I am a foreigner", this communicates something about their role in relation to their current location, which would not be the case if they were instead in their country of origin. Though languages vary in the degree and manner to which such attributes may be communicated, this is a common thread among them, and thus one which we would expect our Natural Language Processing systems to be equipped to handle.

Unfortunately, dealing with the effects of such context on language is still challenging for NLP systems. For example, differences in language identification accuracy, speech recognition word error rates, and translation quality have been observed on the basis of attributes such as a speaker's gender, race, dialect, or role (Blodgett and O'Connor, 2017; Tatman and Kasten, 2017; Tatman, 2017; Stanovsky et al., 2019). In addition, gender has been demonstrated to have an effect on text classification and generation (Prost et al., 2019; Dinan et al., 2020). Moreover, these systems systematically underperform on data generated by those in the minority, having implications for ethics and fairness in regards to the use of such technologies.

Previous work has explored the relationship between user context and language for the purpose of building personalized classification or generation models (Al-Rfou et al., 2016; Dudy et al., 2021; Welch et al., 2022). There is also work that models differences in language usage based on one's register or domain (Sennrich et al., 2016; Yang et al., 2018). However, research that focuses on the effects of intrinsic and extrinsic speaker attributes on language is still limited, mostly dealing with intrinsic speaker attributes in relation to classification tasks (Hovy and Yang, 2021).

In this dissertation, we develop systems that are able to recognize linguistic differences

based on these types of contextual features, centering on both intrinsic and extrinsic speaker attributes. Our contributions include introducing novel datasets for tasks related to such attributes, as well as models to address them. First, we look at the task of gender obfuscation, for which the goal is, given a gendered text, to rephrase the text such that the semantic meaning is preserved while the gender of the writer of the text is masked. We find that there are linguistic differences between female and male authors on the Reddit platform that can be identified by a classification model, and that taking into account the topic of discussion allows for such a classifier to better capture these differences. Moreover, after pairing the classifier with a text generation model, we show that these authors' texts can be rewritten to be more neutral under the classifier while maintaining high similarity to the original text. Then, we explore the task of deception detection, for which the goal is to identify whether or not the writer of a text is taking on a deceptive role. We determine that behavior differs between honest and deceptive players in the Mafia game, and that by using auxiliary tasks that model players' text in context, we are able to discern deceptive actors through their language. We additionally use our approach to reveal features of deceptive language in the game.

The dissertation is organized as follows. In Chapter 2, we describe and analyze aspects of the dataset we use for gender obfuscation, for which we introduce methods in Chapter 3[1]. Chapter 4 centers around the dataset we develop to identify deceptive actors[2], while Chapter 5 presents our approach for using this dataset to detect deceptive language[3]. Finally, we explore future directions and conclude our discussion in Chapter 6.

[1]Chapters 2 and 3 are both adapted from work in submission at the time of writing this dissertation. My co-authors, Risham Sidhu and John DeNero, have both given me permission to include this work as part of my dissertation.

[2]Chapter 4 is adapted from *Investigating the Behavior of Malicious Actors Through the Game of Mafia* (Ibraheem et al., 2020). My co-authors, Vael Gates, John DeNero, and Tom Griffiths, have all given me permission to include this work as part of my dissertation.

[3]Chapter 5 is adapted from *Putting the Con in Context: Identifying Deceptive Actors in the Game of Mafia* (Ibraheem et al., 2022). My co-authors, Gaoyue Zhou and John DeNero, have both given me permission to include this work as part of my dissertation.

Chapter 2

Gender and the Reddit Dataset

We begin our discussion by investigating the effect of an intrinsic speaker attribute, namely gender, on language. In this chapter, we describe how information about a speaker's gender may be gleaned from an online forum platform. We then analyze how gender is expressed in the resulting dataset, demonstrating that linguistic usage varies based on gender. We also show that these gender-based differences are influenced by the subject of discussion, indicating the importance of considering topic when modeling gender and language.

2.1 Introduction

The tremendous recent progress in natural language understanding has largely focused on inferring the meaning or intent of utterances. Language also carries information about the speaker or author's personal attributes, such as their gender. Automatic inference of author gender may leverage various aspects of language, such as gendered lexical items (e.g., "I work as a waitress"), topic choice (e.g., figure skating or hockey), or style/phrasing (e.g., "got my hair done" instead of "got my hair cut").

One challenge in characterizing differences in language usage between genders is with regards to contrast in the distribution of conversed topics. Even gender-neutral text may carry information about author gender through statistical association between topics and gender. For example, one study of online blogs found that the top frequent words by information gain included 'shopping' and 'skirt' for female bloggers, as opposed to 'linux' and 'programming' for male bloggers (Schler et al., 2006), which suggests that a significant source of information about author gender in blog collections is topical rather than performed gender. This highlights the need to account for topic in order to isolate linguistic differences due to gender.

To develop a gender classification system, we annotate a corpus of Reddit comments with user gender by leveraging the fact that corresponding Reddit communities organized around a specific topic exist for users of different genders. We use this dataset to train a BERT-based topic-conditional author gender classifier (Devlin et al., 2018) that is trained to predict the

gender of a user from the set of comments they make about a particular topic or domain (called a *subreddit*). We demonstrate that Reddit comments do indeed carry information about author gender beyond the topic of the text.

To summarize, in this work, we:

1. Label a subset of the Pushshift Reddit dataset for author gender (Baumgartner et al., 2020).
2. Confirm that there are linguistic differences between authors of different genders that can be modeled by a BERT-based classifier.
3. Demonstrate that, by conditioning such a classifier on the topic of discourse, we can capture aspects of gendered language beyond statistical associations between topic and gender.

2.2 Background and Related Work

Personal attributes such as race or gender can contribute significantly to one's identity and may also be highlighted in a person's language usage. In particular, gender can be understood as an attribute that is performed through a variety of acts, which includes the act of writing text (Butler, 1988). Text that exhibits some characteristic of the performance of gender will be described as *gendered text*, whereas a lack of performed gender will be described as *gender-neutral text*. Research has further demonstrated that a myriad of linguistic characteristics, such as lexical items, length of utterances, etc. are correlated with gender (Newman et al., 2008). However, due the performative nature of gender expression, these average tendencies are not an essential characteristic for every individual of a given gender, and prior work has shown that authors may deviate from the stereotypical language usage associated with their gender (Bamman et al., 2014).

Previous research has explored the classification of Reddit users based on gender (Vasilev, 2018; De Pril, 2019). However, gender differences may not only be expressed through linguistic usage, but also may be exhibited through topic selection. Prior work on the Reddit platform has demonstrated differences in terms of both the choice of subreddits as well as the amount of participation within subreddits between male and female users (Thelwall and Stuart, 2019).

In this work, we collect a dataset of Reddit comments and use this data to develop a topic-aware gender classification model, thus allowing for the isolation of gender-specific linguistic differences from those associated with gendered divergence in topic.

2.3 Dataset

For our dataset, we collect comments from the Pushshift Reddit Dataset. Reddit is an online communication platform on which users participate in different communities orga-

Size	***LivingSpace***		***HairAdvice***		***FashionAdvice***	
	Female	*Male*	*Female*	*Male*	*Female*	*Male*
Training	8737	8737	8737	8737	8737	8737
Test	489	489	489	489	489	489

Table 2.1: Number of comments per subreddit pair in the Reddit dataset. For each topic, an equal number of comments were taken from the corresponding female and male subreddits.

nized around shared interests. Redditors are able to create accounts, which they can use to post and comment in multiple such communities, known as subreddits. In order to get gendered texts for our investigation, we searched for subreddit pairs in which the same topic is labelled as either female or male[1]. After filtering for suitability, we were left with the following six subreddits: femalelivingspace, malelivingspace, femalehairadvice, malehairadvice, femalefashionadvice, and malefashionadvice. We assumed that the majority of authors in these subreddits were of the gender indicated by the subreddit name[2], and thus applied this label to all authors in the given subreddit. From these subreddits, we took comments from the one-year period of October 2018 to September 2019 as our training set and used the three-month period of October 2019 to December 2019 as our test set. We ensure that the number of comments for each subreddit is the same so that all genders and topics are equally represented (Table 2.1).

Characterizing Gendered Language

In order to confirm that our data includes examples of linguistic differences between gendered subreddits, we calculated the information gain for each of the words in our training set as follows:

$$-\sum_{o\in\{f,m\}} P(o)\log P(o) + \sum_{i\in\{w,\bar{w}\}} P(i) \sum_{o\in\{f,m\}} P(o|i)\log P(o|i),$$

where i represents the input features, with w indicating that the word appears in the given text and $\bar{w}$ indicating that the word does not appear, and o represents the output labels, which can be either female (f) or male (m).

We then compute a list of the top 20 frequent words by information gain for each gender across all subreddits in order to summarize differences in lexical choice across genders

[1]Though the same method could be used for identifying authors of other genders, gathering a sufficient amount of data for such users may require other sources, which we leave to future work.

[2]Though this introduces some noise into our dataset, using the method for labelling author gender described in Chapter 3 supports our hypothesis that most authors are indeed of the given gender. Specifically, 96.4% of authors in these subreddits that we are able to label with Chapter 3's method during the month of October 2018 are of the given gender.

Rank	*Female*	*Male*
1	!	sides
2	i	man
3	love	barber
4	so	dude
5	cute	beard
6	bangs	guy
7	my	fade
8	and	buzz
9	thank	his
10	color	taper
11	)	chinos
12	’	bro
13	blonde	handsome
14	bob	pomade
15	pixie	he
16	her	hairline
17	have	shave
18	she	bald
19	m	product
20	beautiful	shirt

Table 2.2: Lists of the top 20 frequent words in the training set by information gain for each gender.

(Table 2.2). Looking at each list, we see that the items conform to our expectations of stereotypical language usage (eg. ‘bangs’ for female versus ‘beard’ for male).

2.4 Experiments

Next, we build a classifier to predict the gender of a comment’s author. We use a pre-trained BERT-Tiny model (2 layers, 128 hidden dimension size) and train with a batch size of 32 for 48 epochs (Devlin et al., 2018). We first train the model on the full training set of 52,422 comments and test on the full test set of 2,934 comments, finding that we are able to achieve an accuracy of 75.0% (Table 2.3). Moreover, when training for the same number of steps on only two of the three subreddit pairs, we find that models that were trained on a given subreddit pair perform consistently better when tested on that pair than those that were only trained on the remaining two subreddit pairs, indicating that there are topic-specific differences in language usage between female and male redditors for these

Accuracy	*w/o LivingSpace*	*w/o HairAdvice*	*w/o FashionAdvice*	*All*
LivingSpace	64.2%	68.8%	67.5%	67.4%
HairAdvice	82.1%	71.0%	82.1%	82.1%
FashionAdvice	75.8%	75.1%	66.4%	75.6%
All	74.0%	71.6%	72.0%	75.0%

Table 2.3: Gender classification accuracy on different subsets of the test set for models that were trained on different subsets of the training set. 'w/o' indicates that the model was not trained on the given subreddit pair.

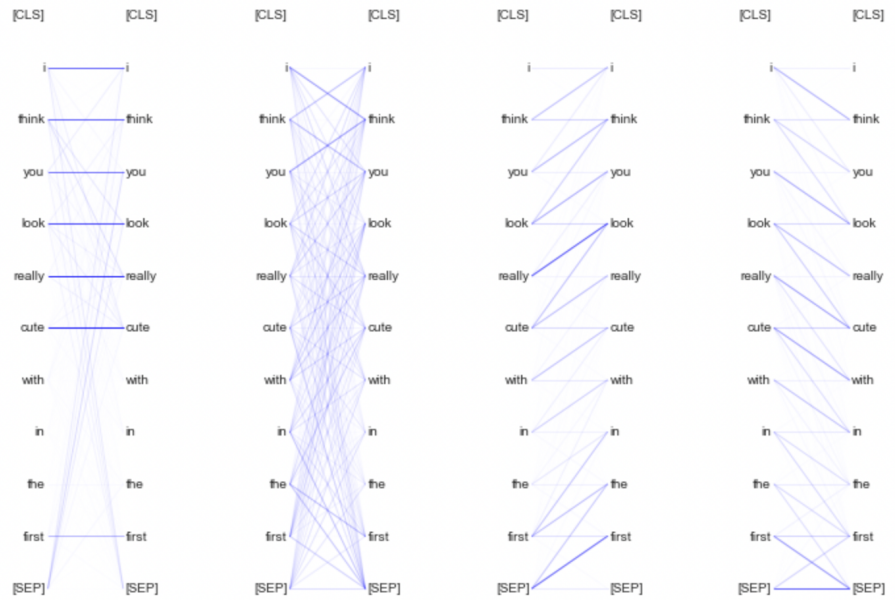
subreddits. Despite this drop in performance, we observe that these models are still able to achieve relatively high accuracy, suggesting that these models are also able to learn aspects of gendered language that are general across topics.

2.5 Discussion

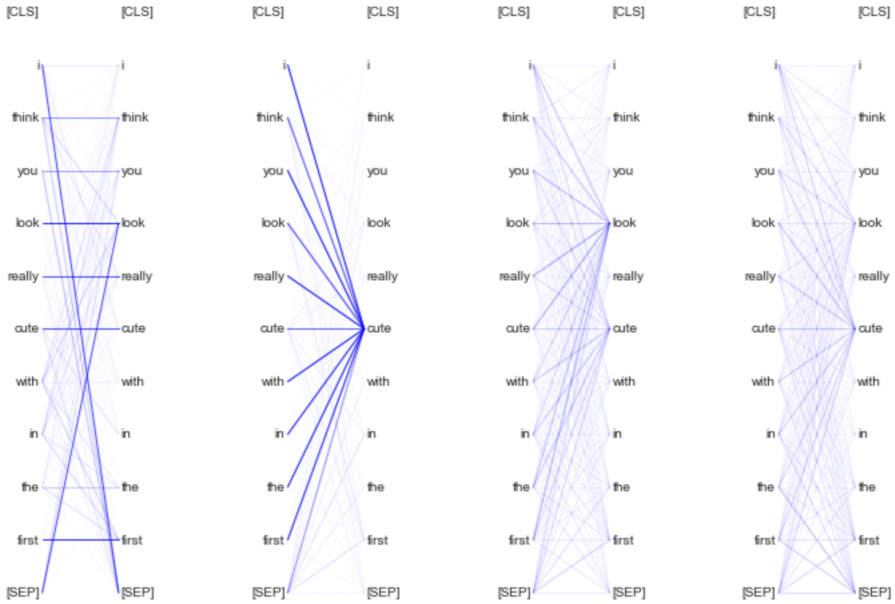
Having trained a model to classify Reddit comments based on gender, we visualize the attentions for some examples to investigate what differences in gendered language it is able to learn (Clark et al., 2019). Looking at a specific female test set example, we find that after finetuning, the classifier assigns more of its attention to a word from our list of top female words by information gain above, further validating that it is able to capture such differences in linguistic usage when making its gender classification decisions (Figure 2.1).

Investigating this phenomenon more deeply, we focus on the second attention head in the first layer of the finetuned model. We calculate the average attention assigned to each word in the test set for female examples that were correctly predicted by the model as female, male examples that were correctly predicted as male, female examples that were incorrectly predicted as male, and male examples that were incorrectly predicted as female. After filtering out words that appear fewer than 3 times in the test set, we compiled a list of the top 20 attended words for correct female, correct male, incorrect female, and incorrect male examples (Table 2.4). From these lists, we can confirm that the model indeed attends to words from our above mutual information lists (eg. 'cute' and 'bangs' for female, 'handsome' and 'beard' for male). Moreover, we observe that more attention is assigned to words from correctly predicted examples than from those that were incorrectly predicted, particularly for female examples, suggesting that the classifier may not detect enough gender information to make an accurate prediction in these cases[3]. Finally, we find similar words when comparing across lists where the model makes the same gender prediction (eg. 'dress'/'dresses' for female, 'sweatshirt' and 'pants' for male), which not only indicates that the model captures

[3]This hypothesis is also bolstered by the fact that the model assigns higher probabilities on average to correct (87.0%) versus incorrect (76.7%) predictions.



(a) Attention weights before finetuning.



(b) Attention weights after finetuning.

Figure 2.1: Visualization of the attention heads for each layer of the (a) pre-trained vs. (b) finetuned BERT model on a female example from the test set. Focusing on layer 1, attention head 2, we see that before finetuning, each word is assigned roughly equal weight **(top)**. After finetuning, the word 'cute', which is of rank 5 on the female list of words by information gain (Table 2.2), receives the bulk of the attention **(bottom)**.

Rank	***Female***		***Male***		***Female***		***Male***	
	Word	*Attn*	*Word*	*Attn*	*Word*	*Attn*	*Word*	*Attn*
1	fairy	0.530	handsome	0.501	socks	0.244	curly	0.420
2	braid	0.496	beard	0.432	sweatshirt	0.240	cozy	0.370
3	necklace	0.476	mate	0.336	hair	0.205	clothes	0.337
4	tattoos	0.475	tho	0.328	jacket	0.197	items	0.311
5	nightstand	0.388	beer	0.282	awkward	0.189	fireplace	0.289
6	dresses	0.359	barber	0.282	wires	0.188	wardrobe	0.284
7	candle	0.351	sweatshirt	0.260	shirt	0.186	sofa	0.274
8	dream	0.347	drivers	0.256	pants	0.186	shopping	0.264
9	lipstick	0.342	tied	0.253	sham	0.183	dress	0.247
10	bangs	0.332	dude	0.237	shorts	0.180	curtains	0.241
11	outfits	0.320	coat	0.228	his	0.178	dresser	0.236
12	cute	0.317	denim	0.226	jeans	0.158	amazon	0.222
13	heels	0.315	shirts	0.219	sofa	0.152	sweater	0.210
14	curling	0.302	chelsea	0.202	women	0.148	women	0.204
15	purse	0.295	bro	0.202	clothes	0.145	awesome	0.194
16	earrings	0.283	slim	0.199	shoes	0.144	gonna	0.162
17	pearls	0.280	pants	0.198	bags	0.135	lamps	0.159
18	bala	0.276	guy	0.196	luxury	0.132	share	0.159
19	blonde	0.265	bomber	0.193	sorry	0.129	interesting	0.157
20	skirt	0.259	man	0.192	leather	0.126	th	0.153

Table 2.4: Lists of the top 20 words in the test set by average attention in layer 1, attention head 2 for each gender on examples where the model correctly (**left**) and incorrectly (**right**) predicted author gender. *Attn* is short for *Attention*.

stereotypical gendered language usage, but also highlights the fact that such language usage is not prescriptive.

2.6 Conclusion

In this work, we show that a classifier may be trained to predict the gender of the reddit user who produced a particular comment within a given subreddit. We demonstrate that taking into account the specific topic of discussion allows us to not only capture gendered language usage that is general across topics, but also such language usage that is unique to a given topic, thus confirming the significance of considering such context in the design of a gender classification system.

Chapter 3

STOPGAP: STyle Obfuscation to Protect the Gender of Authors through Paraphrasing

In the above chapter, we established that language can communicate additional information beyond its meaning, such as the gender of the author, and that neural networks can be used to infer this additional information automatically. An open question is whether neural networks can also be used to prevent this inference and thereby protect the privacy of author attributes. We develop and evaluate methods for *gender obfuscation*, the task of paraphrasing sentences to preserve meaning but reduce or eliminate the amount of information conveyed about the author's gender. Our experiments using Reddit data demonstrate that online text associated with anonymous users does indeed carry information about author gender, and that the data can be used to develop a fully automated gender obfuscation system. We describe a pipeline that includes BERT-based lexical substitution and evaluate the trade-off between faithfulness to the original meaning and the degree of gender obfuscation.

3.1 Introduction

This work explores methods for detecting author gender in text, as well as obfuscating it via paraphrase. A central concern in obfuscating personal attributes of an author while preserving meaning is that some of the information about the author is essential to the meaning of the utterance, while some is largely orthogonal. For example, paraphrasing "I am a waitress" as "I am a server in a restaurant" removes some information about the author gender, but not all, since about 70% of servers identify as female in the United States. But paraphrasing to "I work in a restaurant", which removes nearly all information about author gender, also changes the meaning of the utterance more substantially. Therefore, fully preventing inference about author attributes while fully preserving meaning is often not achievable. However, the indications of author gender that are due to style, lexical

choice, or phrasing can be manipulated while preserving the topic and semantic content of text, potentially reducing the certainty by which author attributes can be inferred while preserving most of the meaning.

To develop a gender obfuscation system, we annotate a larger corpus of Reddit comments with user gender by leveraging Reddit metadata. As in the previous chapter, we use this dataset to train a BERT-based topic-conditional author gender classifier that is trained to predict the gender of a user from the set of comments they make in a particular subreddit. We confirm that these Reddit comments also carry information about author gender beyond the topic of the text, which indicates that at least the topic of text can be preserved, if not the entire meaning, while removing some information about author gender. We then develop and evaluate a paraphrase pipeline that includes BERT-based lexical substitution (Zhou et al., 2019). We show that gender classifier accuracy can be reduced substantially while generating paraphrases that are very similar to the original text.

Our contributions include:

1. A gender-labeled version of a subset of the Pushshift Reddit dataset.

2. An evaluation of a BERT-based author gender classifier that conditions on the topic of discourse in order to model aspects of gendered language beyond statistical associations between topic and gender.

3. A method for applying BERT-based lexical substitution to the task of gender obfuscation and an evaluation of its performance.

3.2 Background and Related Work

Though speakers themselves might take advantage of correlations in linguistic usage to demonstrate their belonging to a specific group, they may also wish to keep this information hidden from their fellow interlocutors. Gender is one such personal characteristic that people might not wish to share online but that can be determined by one's language use. Malicious actors could use information gleaned in this way to identify members of a particular identity for nefarious purposes. For example, one might attempt to discern aspects of a person's identity for the purpose of initiating a spear phishing attack (Brundage et al., 2018). We would therefore like to create natural language processing systems that not only leverage speaker attributes for useful applications, but also allow users the control to protect their identity attributes from others.

The task of obfuscation aims to remove contextual information about the source of a text while maintaining its semantic content. Previous works on obfuscation have looked at obfuscating style through invariance (Emmery et al., 2018), obfuscating authorship through heuristic search (Bevendorff et al., 2019), and obfuscating gender through lexical substitution (Reddy and Knight, 2016).

Obfuscation also has connections to the task of paraphrasing, since both must preserve the semantics of the input. Research in paraphrase generation has explored techniques such as neural machine translation (Mallinson et al., 2017), variational auto-encoders (Roy and Grangier, 2019), iterative search over sequence edits (Liu et al., 2019), and inverse reinforcement learning (Li et al., 2017).

A related task is that of gender rewriting, which aims to rephrase utterances that use gendered lexical items, such as certain pronouns (eg. "she") and profession-related words (eg. "fireman"), in gender-neutral terms (for example, by changing "she" to "they" or "fireman" to "firefighter"). Though previous work has investigated such approaches, these largely use rule-based methods either directly through the rewriting process or indirectly in the data generation process for a neural model, and are thus not designed to capture the gendered linguistic differences that exist outside of the morphological realm of language (Sun et al., 2021; Vanmassenhove et al., 2021; Alhafni et al., 2022).

In order to create a system that provides users agency in hiding aspects of their identity, we rely on advances in the space of natural language generation. While previous generation methods used templates and grammar-based techniques to produce well-formed outputs, limiting the range of utterances that could be produced, more recently statistical methods have enabled contemporary NLP systems to generate language based on patterns in the dataset (Gatt and Krahmer, 2018). This trend has progressed through deep learning, which improved the fluency and coherence of machine-generated text. In addition, these models may be conditioned on some context to control the style of the text that is generated (Keskar et al., 2019).

In this work, we focus on gender obfuscation, with the aim of generating gender-neutral paraphrases of gendered text. This poses an interesting challenge, as though examples of language labeled with author gender are available, labels for gendered versus gender-neutral language are more difficult to collect. Furthermore, using humans to generate or identify gender-neutral examples would likely limit the amount of data available compared to using naturally-occurring sources of text. Unlike previous methods that have either aimed to reinforce or flip the perceived gender of a text writer or used rule-based approaches to generate gender-neutral examples, we thus use a gender classification model to distinguish between gendered and gender-neutral language.

3.3 Dataset

As in the last chapter, we collect comments from the Pushshift Reddit Dataset, this time taking a larger sample of comments from a greater variety of subreddits. Note that subreddits may be very specific, such as 'financialindependence', or quite general, such as 'AskWomen'. For this reason, though we use these subreddits as a proxy for topic, specificity will vary across them. Each subreddit may have a convention for how participants signify aspects of their identity within the community, which are exhibited through their *author flair*. In order to associate text with author gender, we first searched for redditors whose author flair

Author Flair Type	***Search Text***	***Example***	***Subreddits***
CSS Class	pink|blue	pink	tall, short
CSS Class	female|male	male	AskMen, AskWomen, AskMenOver30, AskWomenOver30, sexover30
Text	A/S/L	30/F/JP	OkCupid, keto, childfree, xxketo, LGBTeens, loseit, Tinder, proED, fatlogic, financialindependence, infj, infertility, 100DaysofKeto
Text	♀|♂	♂	All

Table 3.1: List of the methods and corresponding subreddits used to identify author gender for the Reddit Gender dataset. From left to right, the columns indicate: (1) the format in which the author flair is conveyed, (2) the template used to extract gender from author flair, (3) an example author flair that conveys one's gender, and (4) the subreddits that use this format to convey author gender. 'A/S/L' stands for 'Age/Sex/Location'. 'All' indicates that all of the given subreddits are used for this method. An author whose flair matches 'pink', 'female', 'F', and '♀' is labelled as 'female' in the resulting dataset, whereas an author whose flair matches 'blue', 'male', 'M', and '♂' is labelled as 'male'.

could be reliably gendered as either female or male[1] in certain subreddits during the one-month period of October 2018 (Table 3.1)[2]. We then collected comments from the authors whose gender we were able identify that were written across the Reddit platform during the same one-month period. See Table 3.2 for example comments. Finally, we merged all of the comments that were shared by the same author in a given subreddit and kept the subreddits for which there were at least 384 unique authors[3], so that for each subreddit we had 64 authors for both our development and test sets, and at least 256 authors for our training set. This left us with 106 subreddits total. Table 3.3 contains dataset statistics. The training set has 641,353 comments in total, with 448,945 of these comments used to train our models and the remainder set aside for validation purposes. Our development and test sets consisted of 40,388 and 42,456 total comments, respectively.

Characterizing Gendered Language

In order to understand what linguistic differences are associated with gender in our dataset, for each subreddit in our training set, we calculated the information gain of each word as follows:

[1]Though the same method could be used for identifying authors of other genders, gathering a sufficient amount of data for such users may require other sources, which we leave to future work.

[2]This approach is modelled after `https://github.com/bburky/subredditgenderratios`.

[3]Note that we allow overlap of authors across different subreddits.

Subreddit	***Author***	***Gender***	***Comment***
OkCupid	cool_username	female	Right?! Something straight from hell
OkCupid	username_tbd	male	It's literally 2 dimensional characters walking into a room and ranting about ideology for 700 pages.
bestof	cool_username	female	Yup, it's the Jesus Bar if it's in the dash.
bestof	username_tbd	male	won't someone think of the poor, oppressed wealthy people.

Table 3.2: Dataset examples taken from the dev set, where usernames have been anonymized. Note that the dataset includes both female and male authors in the same subreddits. Also observe that we include both subreddits that were used to collect author gender information as described in Table 3.1 (eg. OkCupid) as well as those to which gender information was propagated through the author (eg. bestof).

	Female	***Male***	***Total***
Total number of authors	8312	10404	18716
Total number of comments	243634	397719	641353
Total number of SA pairs	33204	56121	89325
Total number of SAC tuples	240013	394712	634725
Avg number of comments per SA pair	7.337	7.087	7.180
Std number of comments per SA pair	29.475	29.728	29.634
Avg length of comments per SA pair	1873.583	1305.411	1516.611
Std length of comments per SA pair	11553.085	5993.173	8500.333

Table 3.3: Statistics for the Reddit Gender training dataset. *SA* is short for *Subreddit-Author*, *SAC* is short for *Subreddit-Author-Comment*, *Avg* is short for *Average*, and *Std* is short for *Standard Deviation*. There is no statistically significant difference between female and male authors in the average number of comments per subreddit-author pair ($p = 0.22$). The difference bewteen female and male authors in the average length of comments per subreddit-author pair is statistically significant ($p < 0.0001$)

$$- \sum_{o \in \{f,m\}} P(o) \log P(o)$$
$$+ \sum_{i \in \{w,\bar{w}\}} P(i) \sum_{o \in \{f,m\}} P(o|i) \log P(o|i),$$

where i represents the input features, with w indicating that the word appears in the given text and $\bar{w}$ indicating that the word does not appear, and o represents the output labels,

Rank	*Female*	*Male*
1	husband	f*ck
2	boyfriend	thats
3	omg	way
4	😂	wife
5	makeup	yea
6	lovely	im
7	bc	reddit
8	pregnant	likely
9	grateful	government
10	wedding	american
11	haven	fire
12	meds	dont
13	😊	gun
14	bf	comments
15	pregnancy	good
16	apartment	against
17	abusive	mind
18	shopping	tax
19	adorable	money
20	pets	guess

Table 3.4: Lists by gender of the 20 most frequent words across top-1000 word lists by information gain for each subreddit in the training set. '*' indicates censorship of a pejorative.

which can be either female (f) or male (m).

We then computed a list of the top 1000 words by information gain for each gender and subreddit in order to summarize differences in lexical choice across genders. After filtering out URLs and words that did not appear in at least one-tenth of the lists for each gender (ie. for at least 11 different subreddits), as well as words that appeared on the lists for both genders, we finally compiled a list of the top 20 words for each gender according to how frequently each word occurred across the subreddit lists (Table 3.4).

Looking at each list, we see that the items include some examples of stereotypical language usage across genders (eg. emojis for female versus pejoratives for male), as well as some examples of stereotypical topics discussed by each (eg. 'shopping' versus 'government'). There are also stylistic differences: males appear to avoid apostrophes ('thats', 'im', 'dont') while females appear to use abbreviations ('omg', 'bc', 'bf'). These patterns indicate that gender manifests itself in a variety of ways within the dataset.

Method	***Train Acc***	***Dev Acc***
Majority	0.628	0.618
Comment	0.721	0.718
Author	0.722	0.723

Table 3.5: Gender classification accuracy for three baseline methods.

3.4 Approach

Gender Classification

Since our dataset contains not only authors' genders and comments, but also topic information by way of subreddit, we first compute the accuracy for several baseline methods that make use of the gender distributions of these subreddits. For our first baseline, we calculate the majority gender for each subreddit based on the number of comments written by female and male authors in the training set. We then classify all dev set subreddit-author examples with the majority gender for the corresponding subreddit using this method. For our second baseline, we find the majority gender for each subreddit based on the number of female and male authors in the training set. We then classify all dev set subreddit-author examples with the majority gender for the corresponding subreddit using this second approach. As shown in Table 3.5, the baselines are able to achieve accuracies better than predicting the majority label over all of the training data, indicating that author gender ratio is statistically associated with subreddit topics.

Next, we build a gender classifier for our data. We consider two main cases: where we have a single comment from an author and where we have all of an author's comments in a single subreddit. Generally, the case with only a single comment is much more difficult as an individual comment may not contain that much information about a speaker's gender, especially if it is short. Within both of these cases, since we found that subreddit information was useful in classifying gender in the baselines, we added a token to the beginning of each example to represent the subreddit. We use a pre-trained BERT-Tiny model (2 layers, 128 hidden dimension size) as larger models quickly overfitted to the data and lead to large performance gaps on unseen comments, authors, and subreddits. We train with a batch size of 8 for 4 epochs. Results are shown in Table 3.6.

By adding an additional linear layer on the output of the Stacked + Topic model, adding an intermediate dropout layer, and training for 6 epochs, we are able to achieve an accuracy of 78.9% on the training set and 75.2% on the dev set. Moreover, we are still able to achieve high accuracy when training for the same number of steps on only a subset of the 106 subreddits, suggesting that the model is able to learn aspects of gendered language that are general across topics.

We also investigated whether the subreddit affected the interpretation of gendered language as follows. We replaced each topic token with one of a subreddit that had a similar

Method	***Train Acc***	***Dev Acc***
Single - Topic	0.666	0.702
Stacked - Topic	0.737	0.704
Single + Topic	0.710	0.743
Stacked + Topic	0.764	0.716

Table 3.6: Gender classification accuracy on train and dev sets for a BERT-based model trained on individual comments (Single) vs. all of an author's comments in a subreddit (Stacked), with explicit topic representations (+ Topic) or without (- Topic)

gender ratio to provide the same topic-based statistical association to the model but break the link between topic and word choice. This substitution caused a small drop in performance (1-2%), suggesting that the model is learning a topic-specific interpretation of gendered language in some cases. This highlights the importance of modeling the topic and text jointly.

Gender Obfuscation

In order to build our gender obfuscation system, we develop a pipeline that identifies a gender-indicating phrase, generates candidate replacements, and selects a replacement according to an obfuscation objective. We describe three variants, a fixed-length method, a variable-length method, and a lexical substitution method. We use our gender classification model to rank gender-neutral substitutions, ie. those that assign the text equal probability of being predicted as either female or male, over gendered ones. For example, given the comment "You look beautiful", we would expect 'magnificent' to be rated above 'handsome' or 'stunning' as a possible substitution for 'beautiful'.

Fixed-Length Masking

For our simplest method, we identify what to paraphrase by using our gender prediction model to compare the gender prediction differences when each word or phrase is masked out of the input sentence. The phrase span p whose masking causes the most difference between the original gender prediction and the new one is the most salient indicator of gender, and we select this for replacement:

$$p = \text{argmax}_{p \in N_3} |G(x) - G(\text{mask}(x, p))|,$$

where G is the gender prediction model, x is the original text, $\text{mask}(x, p)$ is the original text with the n-gram p replaced with `[MASK]` tokens, and N_3 is the set of all possible n-grams of length up to 3 in the text (i.e. all uni-, bi-, and tri-grams). We suggest alternate words and phrases using the standard masked token prediction task used to train a SpanBERT model, which is a BERT variant that performs its pre-training MLM task on spans of tokens rather than individual tokens for better sequence understanding (Joshi et al., 2020). A

SpanBERT model fine-tuned on our dataset predicts the top 5 tokens for each masked token in the phrase. Finally, we rank the potential paraphrases by evaluating combinations of these tokens using G. The combination which produces the most neutral prediction is selected:

$$x' = \text{argmin}_{p' \in \text{gen}(x,p)} |G(\text{rep}(x, p, p')) - 0.5|,$$

where $\text{gen}(x, p)$ is the top phrases predicted by the SpanBERT model and $\text{rep}(x, p, p')$ is a paraphrase of x that replaces p by p'.

Variable-Length Masking

When qualitatively evaluating the results of the fixed-length model, we saw that requiring a replacement for an n-gram to preserve the length n restricted the possible replacements. Consider the case: "I feel like young adult fiction is looked down upon.". The phrase "I feel like" is associated with female authorship and the replacements might be: "I think that", "I feel that", etc. But this restraint precludes replacements such as "It seems" or just deleting the phrase altogether and leaving "Young adult fiction...". To allow for such possibilities, we introduce a variable-length version of this approach.

For the variable-length approach, we preserve the method for selecting which phrases should be replaced as well as the final ranking based on the maximal gender prediction change. However, when suggesting replacements for the phrase, we allow for the replacement to be a different length than the original phrase. To ensure that changing this token length does not introduce disfluencies, we employ a *goodness predictor* model that scores whether replacing the phrase constructs a well-formed sentence. This predictor takes in a text with two special tokens marking the start and end of the replaced phrase. It is trained by placing these tokens randomly around n-grams with $n \in [0, 3]$ to represent correct spans and placing them around random deletions and insertions to synthesize disfluent paraphrases. Using this goodness predictor, we perform a preliminary ranking to choose the top 5 suggestions before passing these in to the gender predictor, which selects the one that obfuscates gender the most.

Lexical Substitution

Neither the fixed- or variable-length masking methods preserve the meaning of the phrase that is replaced; their only objective is obfuscating the gender, so some replacements change the meaning of the sentence. For example, in "My girlfriend and I got married yesterday!", replacing "girlfriend" with "boyfriend"/"husband" alters the meaning, whereas "partner"/"spouse" fits better semantically. To address this, we propose using lexical substitution.

BERT-based lexical substitution is a method for replacing individual words in a sentence with their most appropriate synonyms. Such synonyms should not only reflect meaning of the replaced word, but also that of the sentence. For example, when replacing the word "hard" in "The homework is hard", we prefer words like "tough" or "difficult" that encapsulate the intended meaning over words like "easy" (opposite) or "done" (unrelated). To best consider

Obfuscation Pipeline	***Cosine Similarity***	***NDD***	***Gender Prediction***
No Change	1.000	0.000	0.608
Fixed-Length	0.984	0.093	0.576
Variable-Length	0.986	0.085	0.580
Lexical Substitution	0.994	0.063	0.574

Table 3.7: Quantitative results for the three obfuscation pipelines. *NDD* is short for *Neutral Distance Difference*.

both the local context and the global context, the target word's embedding has dropout applied to it (rather than completely masking it out) and a sequence encoding is produced. Then, the predicted distribution for the masked word is balanced against the overall sequence similarity, to pick the best replacement as follows:

$$s(x'_k|x,k) = s_v(x'_k|x,k) + \alpha \times s_p(x'_k|x,k),$$

where x'_k is the candidate substitution for the original token in x at position k, s_v considers global similarity by weighing the cosine similarity of all tokens in the original and dropout-masked contextualized outputs by the selected token at k's attention to each other token, s_p considers local similarity by normalizing the dropout-masked token's distribution over all vocabulary tokens but the original one, using a token's probability to measure similarity to the original token at k, and α is a parameter used to weigh the relative importance of s_v and s_p in calculating the overall similarity $s(x'_k|x,k)$.

We modify this to also include gender and apply it over all the tokens in the n-gram span. Retaining the method of selecting a phrase to be replaced from the base obfuscation, the selected phrase is fed into lexical substitution to choose a paraphrase preserving sequence meaning and token meaning, and neutralizing the gender most:

$$m(x'_k|x,k) = \beta \times |G(x'_p) - 0.5| + \sum_{k \in p} s(x'_k|x,k),$$

where β is a parameter used to weigh the relative importance of the gender-neutralization and similarity terms. We preserve the original weights for lexical substition (1 for the overall preservation, alpha = 0.01 for the local preservation) and use beta = 15 for the gender difference. The word that ranks highest on this metric is selected as the replacement.

3.5 Experiments

The metrics we use to compare our gender obfuscation methods are:

1. the cosine similarity of the original and obfuscated sentences,

2. the difference between the distances of each sentence to a neutral prediction, what we refer to as the *neutral distance difference*, and

3. the gender prediction of the obfuscated sentence.

The first metric measures the semantic similarity, that is, how much of the original meaning is preserved, the second how much the perceived gender is changed, and the third how gender-neutral the new sentence is.

For the cosine similarity metric, 1.0 naturally means that the two sentences encode the exact same meaning, but for other values it is be unclear how much meaning is preserved. To produce a lower bound, we compared completely unrelated sentences and found that their similarity is 0.932 on average. The relatively high number is likely due to the sentences not being complete opposites and using similar syntactic structures included in the sequence representation. Having established a lower-bound, we desired a meaning preservation bound indicating sentences with essentially the same meaning. Since our methods are quite targeted, the overall structure of the sentences used to compute this needed to be quite similar, or our methods' score would be inflated from preserving most of the sentence. We used the JFLEG dataset meant for correcting incorrect human sentences and took the set of ground truth sentences per example as sentences that preserved meaning and structure (Napoles et al., 2017). This gave us a reasonable numerical threshold for preserving information at 0.994 on the scale. Two sentences with scores near this mark convey almost entirely the same meaning.

For the neutral distance difference, we are comparing how much closer a paraphrased sentence is to a neutral gender prediction than its original:

$$\text{NDD}(x, x') = |G(x) - 0.5| - |G(x') - 0.5|.$$

As most predictions are not at the extreme ends of the spectrum ($[0, 1]$), a change of 0.1-0.25 is often enough to make a gendered text fairly neutral, and given our prediction range, this difference falls in $[-0.5, 0.5]$, where a negative number indicates a more gendered paraphrase.

As our goal is to obfuscate the gender, we want the paraphrase's gender prediction to be as close to 0.5 as possible, with the proximity indicating the strength of the method. The results for our gender obfuscation approaches using these metrics may be viewed in Table 3.7.

3.6 Discussion

We see from Table 3.7 that the fixed-length masking method is able to affect the gender prediction the most, as it only considers neutralizing the prediction (shown by the lower cosine score indicating some lost meaning). The variable-length method gives a slight increase in cosine similarity since its goodness predictor is filtering out paraphrases that don't match the structure of the sentence, but so loses replacements that might neutralize gender more. The lexical substitution method produces the the least change in gender prediction, but preserves the overall sentence meaning the most.

Original Text	***Fixed-Length***	***Variable-Length***	***Lexical Substitution***
1. I like the author, the cover is aesthetically pleasing and recs from friends	*I like* ***that that*** *the cover is aesthetically pleasing and recs from friends*	*I like* ***this because*** *the cover is aesthetically pleasing and recs from friends*	*I like the author, the cover is aesthetically pleasing and recs from friends*
2. Not trying to be rude but how did you get to this age not knowing this?	*Not trying to be* ***cool and when*** *did you get to this age not knowing this?*	*Not trying to be* ***cool and why*** *did you get to this age not knowing this?*	*Not trying to be* ***inappropriate but how*** *did you get to this age not knowing this?*
3. It's two years age difference how tf is that grooming?	*It's two years age* ***...*** *is that grooming?*	*It's two years age* ***.. what*** *is that grooming?*	*It's two years age* ***. how tf*** *is that grooming?*
4. Hey! That's insult to us 20 yr olds :(	*Hey!* ***so listen*** *to us 20 yr olds :(*	*Hey!* ***so look*** *to us 20 yr olds :(*	*Hey!* ***That's rude*** *to us 20 yr olds :(*
5. the comments, bruh	*the comments,* ***:***	*the comments,* ***: ****	*the comments,* ***briki***
6. you must be so cute and dainty,,, i wish i could be that pretty	*you must be so* ***big ##y but*** *i wish i could be that pretty*	*you must be so* ***beautiful ..*** *i wish i could be that pretty*	*you must be so* ***nice and dainty*** *i wish i could be that pretty*

Table 3.8: Qualitative examples of the three obfuscation pipelines. Changed spans are in **bold and underlined**.

These trade-offs between gender neutrality and meaning preservation and grammaticality are demonstrated in the selection of handpicked examples in Table 3.8. The fixed-length method, which only aims to reduce the gender prediction, picks up on arbitrary patterns and tokens that do not fit together semantically or syntactically. This is visible in the first example with "that that" and the final example where the token '##y' is predicted without an appropriate previous token to attach to. The fixed-length method often ignores the original meaning entirely, changing "that's insult[ing]" to "so listen" and "rude but how" to "cool and when".

The variable-length method addresses one of these issues by filtering suggestions with a goodness predictor and allowing variable-length spans. This is seen in the last example, where rather than a hanging word-end token, the tri-gram is replaced by the single word 'beautiful'. It also avoids repetitions like "that that" (which often appear when the model prefers a shorter span suggestion). But, it is not able to address the issue of preserving the

original meaning, still replacing 'rude' with 'cool' in the second example.

By considering the original context, the lexical substitution method often keeps the original structure, which helps avoid spurious signals learned by the gender prediction model, as weak connections cannot overpower the weight on preserving the local and global context. Paraphrases are closer to the meaning of the original sentence: replacing 'rude' with 'inappropriate' is a better choice than 'cool' (ex. 2), and replacing 'insult[ing]' with 'rude' (ex. 4) preserves the meaning and is more grammatical than the original. However, since this method is a token-to-token replacement method, the paraphrases are of the same length as the selected span and each token plays the same or similar roles to the original constituents. So, it is unable to replace the span "cute and dainty" (ex. 6) with 'beautiful' as the variable-length pipeline does.

The fifth example is interesting due to the term 'bruh', which is likely unfamiliar as it's slang and is perceived as performing masculinity. The first two methods attempt to address this by creating an emoticon. The fixed-length method only provides the colon (highly female coded from its presence in most emoticons), while the variable-length method adds an asterisk (not an emoticon we know, but matching the general form).

Also of interest is the third example, where 'tf' (abbreviation for "the f*ck") is male-coded and thus selected for replacement with nearby words. The fixed-length method replaces the entire phrase with ellipses, changing the tone of the comment from an aggressive attack to a gentler question. The lexical substitution method, trying to preserve the original context, keeps "how tf" as a phrase, but takes out 'difference'. 'Difference' being selected for paraphrasing highlights weaknesses with the selection method not respecting the internal structures of the sentence. Selections like "the author" or "cute and dainty" work well, but "difference how tf" clearly has constituents of different phrases that make it harder to produce meaningful replacements. The final example also highlights how our span selection method doesn't necessarily select the most salient span to a human: rather than the phrase "cute and dainty", what indicates the gender is "i wish i could be that pretty" and a replacement to the last word would obfuscate the gender more.

The provided examples demonstrate the promise of using gender-aware models for the purpose of obfuscation when large amounts of parallel data are not available[4]. Modeling such author attributes may benefit NLP systems more generally, as an important ethical challenge in the field is that performance differences have been observed on the basis of intrinsic and extrinsic speaker attributes. While our gender obfuscation work does not address this challenge directly, the dataset we created is relevant to this challenge, as are the methods we describe for author attribute inference and obfuscation.

[4]We attempted obfuscating the given examples through prompts to GPT-3, as well as by providing a small set of parallel obfuscation pairs, but observed the tendency to misunderstand the task or to overfit to the samples, suggesting that parallel data may be more useful for such an obfuscation approach.

3.7 Conclusion

A person's language depends not only on the content that they wish to convey, but also on the context within which they convey it. Intrinsic speaker attributes such as gender contribute to such context. In this work, we leverage this attribute in order to address the task of gender obfuscation. Though this problem is far from solved, we hope that our efforts help to shed light on how these aspects of context contribute to linguistic usage and provide insights into how to provide more privacy and agency to users in our increasingly virtual world.

Chapter 4

Role and the Mafia Dataset

In the previous two chapters, we presented data and methods for modeling gender, an intrinsic speaker attribute. In this chapter, we pivot to the extrinsic speaker attribute of conversational role, focusing specifically on the role of deceptive actors in a network. Large-scale deception is an interesting game: deceivers must find ways to draw unknowing group members to their side, and group members must develop strategies for detecting deception. What are the strategies that people take on in these roles, and are these behaviors distinct enough that computer systems as well as peers could detect them? To address this question, we analyze how people play in text-based games of Mafia, wherein players are assigned to deceptive roles (mafia) or roles incentivizing detecting deception (bystanders). We find that participants adopt sophisticated role-based strategies, where the mafia, who are outnumbered but know the identities of all players, act carefully to secure the votes of the bystanders by speaking more, even as verbose speakers tended to be eliminated. Additionally, these role-based behaviors were distinct enough that a computational classifier could distinguish between mafia and bystanders with 70.3% accuracy and outperform human players. Understanding the implicit strategies and systematic features that define participants as deceptive or honest advances our ability to automatically detect deceit in online group contexts, and hints at the complexity of group dynamics and non-linguistic cues present in real-world collaboration.

4.1 Introduction

Humans are a largely collaborative species: strangers on the internet exchange goods and services, online social groups form and thrive, people work remotely and communicate on online platforms. However, people sometimes have goals that incentivize them to deceive others. People may bend the truth during negotiations, and in a world that is becoming increasingly virtual, people may not know what conversations are with deceptive or malicious entities without the aid of normal audiovisual cues. Understanding what cues and interaction styles people adopt when behaving deceptively, or when seeking to detect deceptive behavior,

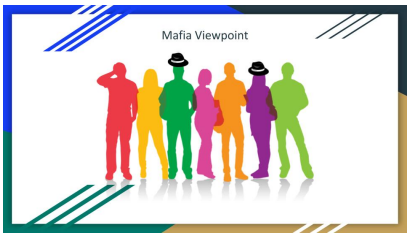
will be crucial to developing both automated detection and a greater understanding of the complex interactions that people use in enacting and revealing deception. Previous work indicates that people struggle with telling apart lies from truth, especially with deceptive statements (Bond Jr and DePaulo, 2006). This raises the question of what strategies deceptive actors use to avoid detection, as well as what strategies honest actors use to discover deceivers.

Deception is a hard topic to study, however, because of its inherent complexity: multiple people with different motivations are trying to evaluate one another, while contending with moral obligations and accusations, over a period of time that involves planning, taking actions, and responding to others' actions. Worse, these calculations and modeling of other people happen in real time. Moreover, there is a distinction between a falsehood, which is a statement that is not true, a lie, which is a statement that the speaker does not believe, and deception, which is the act of convincing another person to hold a false belief. Whereas falsehoods and lies are properties of statements, deceptive intent is a characteristic of the speaker. Therefore, though deceptive speakers may tell falsehoods and lies, they might also provide truthful statements, and vice versa for honest speakers, thus rendering the truth conditions of individual utterances as unreliable indicators of deception. We are interested in how people solve the dual problems of deceiving and detecting deception, which requires a paradigm wherein we can observe all agents' actions and communication while simultaneously knowing agents' incentives and goals. We thus turn to a game with a rich history of deception research: Mafia.

Specifically, we investigate how people determine deceptive actors by engaging participants in an online, text-based game of Mafia (Figure 4.1). In the game, participants are divided into two groups, the mafia (deceivers) and the bystanders (deception-detectors). The mafia know who is a mafioso and who is a bystander, but the bystanders do not. The game is won when either the mafia outnumber the bystanders or the bystanders have eliminated all of the mafia. In each round of the game, during the nighttime phase the mafia have a chance to discretely discuss and eliminate a player, and in the daytime phase all players (the smaller mafia group and the larger group of bystanders) can publicly talk and choose who to eliminate. Since the bystanders are aware of the existence of a deceptive party, the mafia must blend in to evade elimination. On the other hand, the mafia must influence the bystanders, convincing them to eliminate other bystanders, in order to achieve their goals.

In our Mafia game participants interact via text only, rather than adding the complexity of audio and visual cues. Within this medium, participants could freely communicate with each other to discuss their suspicions, question each other, and determine who to eliminate as a group. Participants' responses and voting patterns gave us a rich dataset, allowing us to examine what features governed players' interaction isolated from the intertwined complexities of facial, body, and auditory information normally present in Mafia games. This paradigm let us study the behavior of people with authentic incentives in a scenario with an open-ended but computationally-manageable set of actions, while maintaining reliable annotations of who was being deceptive and who was not.

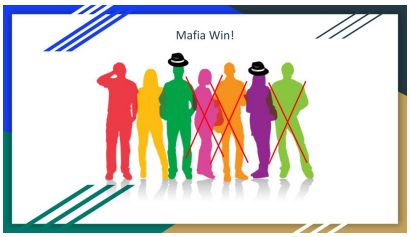
Below, we examine how mafioso and bystander behavior differs empirically, finding that



(a) Mafia members are aware of the identities of other mafia members.

(b) Bystanders are unaware of other participants' allegiances.

(c) The goal of the mafia is to eliminate bystanders until they outnumber them.



(d) Example of mafia achieving their goal as described in (c).

(e) The goal of the bystanders is to eliminate all mafia members.

(f) Example of bystanders achieving their goal as described in (e).

(g) During the nighttime, mafia privately communicate and vote.

(h) During the daytime, everyone publicly communicates and votes.

Figure 4.1: Illustration of the Mafia game. Hats are used to denote mafia members.

mafia members spoke frequently to elicit collaboration from the bystanders, but also acted to avoid the compensating mechanism that all group members punished those who speak too

much. Additionally, we found that players acting as mafia and bystanders behaved distinctly enough that we could create a classifier to differentiate between mafia and bystanders solely based on surface features about players' behavior.

The implications for this work are two-fold. The first is in a computer security context. The patterns of incentives in our paradigm are very similar to the real-life scenario of an organization that knows there is a security breach, but does not know who is responsible. Understanding the strategies adversaries may use to avoid detection in such a network could have implications for general security systems, since developing automated mechanisms of detecting malicious parties could protect users. The second application is psychological. To understand how people engage in deception and detection of deception in group contexts, we need to observe people's behavior when they choose to deceive. We conducted an empirical investigation in a controlled context, wherein individuals were randomly assigned to play a given role of deceiver or deceit-detector, to determine what strategies players use and what features of their behavior could be generalized.

4.2 Background and Related Work

The game of Mafia is particularly well-suited for the goal of determining whether the deceptive participants in a conversation can be identified from the contents of their utterances.

Deception in Language

The idea that deception can be detected based on linguistic cues is intriguing, and previous work has explored scenarios that either mimic or are taken directly from real-world investigations of potentially deceptive actors. Derrick et al. 2013 showed that deceptive parties take longer to formulate responses and use fewer words in the context of chat-based communication. Burgoon et al. 2003 similarly found that deceivers sent briefer chat messages. Fuller et al. 2011 demonstrated the effectiveness of training classifiers to identify deceptive language in relation to crimes, and found that word quantity was a particularly useful feature. Fornaciari and Poesio 2013 also found surface-level features useful in detecting deceptive statements in a criminal context, specifically through the investigation of Italian court documents, while Mihalcea et al. 2013 found that written lies were easier to detect than transcripts of spoken ones. Abouelenien et al. 2014 took a multimodal approach to deception detection, making use of acoustic, thermal, and physiological information to discern liars, and found that non-contact approaches were able to match or exceed the performance of those that were more invasive.

The Game of Mafia

Researchers have also examined deception in games, focusing on settings such as Diplomacy or negotiation over a set of items (Lewis et al., 2017; Niculae et al., 2015). In addition,

	Mafia	Bystander	Total
Total number of players	96	364	460
Average number of players per game	1.96	7.43	9.39

Table 4.1: Mafia dataset statistics. **Mafia** and **Bystander** denote the mafia and bystander classes, respectively, while **Total** denotes the total number for both groups.

there has been some work exploring the effects of biased voting on group decision making (Kearns et al., 2009). The game of Mafia specifically has attracted attention, and researchers have analyzed data from various online game communities. Zhou and Sung 2008 discovered differences between deception across cultural communities by analyzing data from an online Chinese Mafia game, Pak and Zhou 2011 used social network analysis to detect deceivers using the epicmafia.com website, and de Ruiter and Kachergis 2018 collected and trained models on a dataset from the online Mafiascum forum. Researchers have also studied the game of Werewolf, a variant of Mafia. Chittaranjan and Hung 2010 used audio information to classify deceptive parties, while Demyanov et al. 2015 used video information. Braverman et al. (2008) and Migdał (2010) developed a mathematical model of the Mafia game, assuming that all votes are cast at random, which allowed them to analyze how mafia and bystander win rates varied with role distribution in a highly controlled version of the game. Bi and Tanaka 2016 showed that under certain conditions, the strategy of mafia pretending to be bystanders is suboptimal.

Most of the deception-oriented games that have been studied provided individual incentives to the players. We were interested in the Mafia game because it focuses on how patterns of deception arise when incentives are only at the group level. In addition, whereas using datasets of online Mafia games presents a rich source of deceptive language, the complicated nature of games on these forums makes it challenging to isolate specific strategies that participants use to engage in and detect deceptive behavior. In contrast to work using video or audio, we assume that players do not have access to any audiovisual clues about others' identities, thus proposing a more stringent threat-detection model, which we believe is more congruent with the majority of interactions that users have with unverified parties online. Finally, though analyzing mathematical models of Mafia gives insight into certain game mechanics, studies like this ignore the strategies that actual players use in order to conceal their own or discover others' identities. This work takes these factors into account by allowing for a controlled environment that nonetheless supports the use of complex strategies for deceiving and detecting deceptive behavior.

4.3 Dataset

Participants

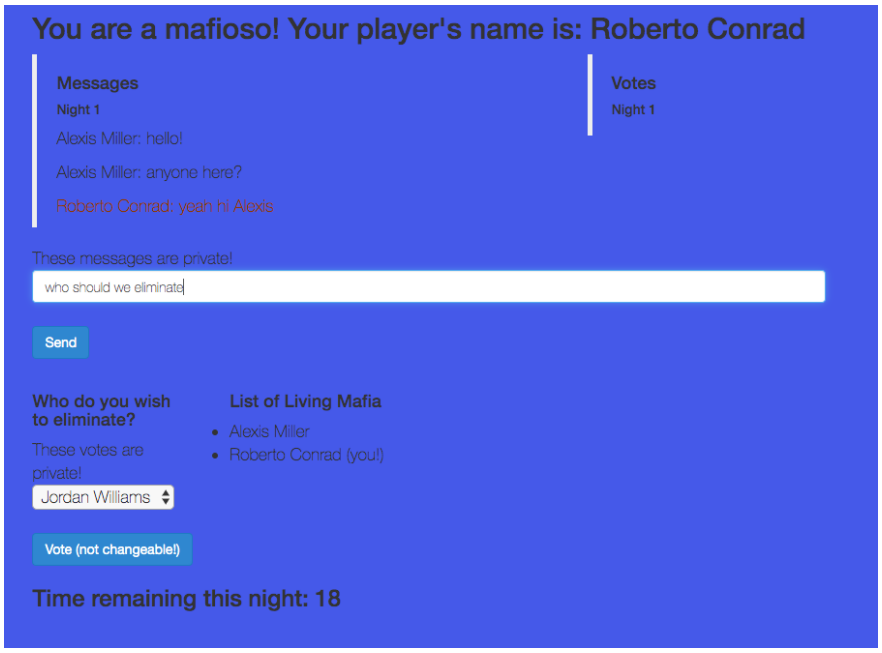
A total of 460 English-speaking participants based in the United States were recruited from Amazon Mechanical Turk using the experiment platform Dallinger[1]. Between 4 and 10 participants were recruited for each Mafia game: 1 to 2 participants were designated mafia, and the rest were bystanders (Table 4.1). Forty-nine Mafia games, which were played during the period from November 2018 to March 2019, are included in the analysis. Participants were paid $2.50 for completing the task, which took about 12 minutes on average, plus bonuses for time spent waiting for other participants to arrive in a chatroom to begin the experiment. Waiting was paid at $5/hour.

Procedure

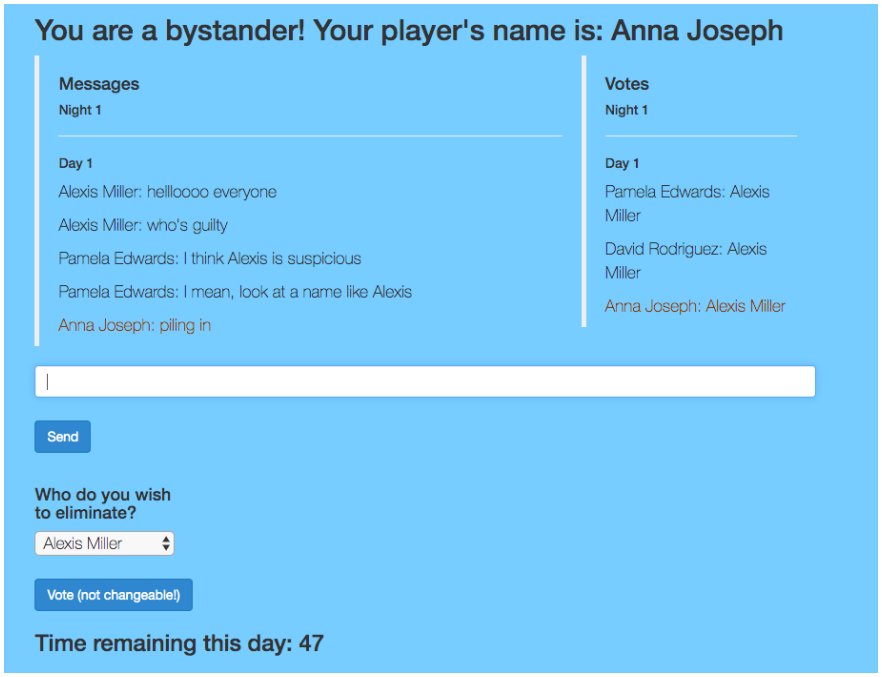
Upon recruitment, participants were shown a consent form, per IRB approval, followed by an instructional video and accompanying transcript describing how to play the text-based Mafia game using an interface we developed (please see Appendix A). After they completed a quiz demonstrating they understood the information, they entered a waiting room until the desired number of participants was reached. Participants were then assigned a role (mafioso or bystander) and fake name, after which they began playing the game.

The game dynamics were as follows. Each mafia member was aware of the roles of their fellow mafia members and thus, by process of elimination, knew the roles of the bystanders. However, the bystanders did not know the true role of anyone else in the game. The goal of the mafia was to eliminate bystanders until the number of mafia was greater than or equal to that of the bystanders. The goal of the bystanders was to identify and eliminate all of the mafia members. Since the incentive structure was set up such that bystanders benefited from true beliefs about who the mafia members were, whereas mafia members benefited from false beliefs, bystanders were thus motivated to be honest actors, whereas mafia members were motivated to be deceptive actors in the Mafia game. The game proceeded in phases, alternating between nighttime and daytime (Figure 4.2). During the nighttime, mafia members could secretly communicate to decide on who to eliminate, after which they discretely voted, and the person with the majority vote was eliminated from the game. If there was a tie, one of the people involved in the tie was randomly chosen to be eliminated. During the daytime, everyone was made aware of who was eliminated during the nighttime, and then all players could openly communicate to decide who to eliminate. All the players then voted publicly, and the person with the majority vote was eliminated and announced to be a bystander or mafioso. Thus, during the nighttime mafia could secretly communicate and eliminate anyone, whereas during the daytime mafia could participate in the voting and communication protocols in the same way as bystanders. The game proceeded until there was a winning faction according to the goals described above.

[1] `http://github.com/dallinger/Dallinger`



(a) Nighttime mafioso view.



(b) Daytime bystander view.

Figure 4.2: Mafia experiment screenshots during **(top)** the first nighttime phase, with the participant as a mafioso, and **(bottom)** the first daytime phase, with the participant as a bystander (note that mafia messages are not visible).

With regards to the voting procedure, we allowed players to vote for anyone except for themselves, as well as to decide not to vote. This decision was made to allow for increased flexibility in the strategies used by participants for deception and deception detection. For example, a mafia member could choose to vote for another mafia member, or to not vote for anyone at all, in order to evade suspicion from the bystanders.

From these experiments, we collected a dataset consisting of both mafia and bystander utterances over the course of each game, as well as the participants' voting behavior. Figure 4.3 displays a snippet of the daytime dialog from one Mafia game. As shown, many utterances are either social interactions (eg. "hi erybody") or discussions about what to do in the game, such as accusations or comments about voting (eg. "I bet it's Mandy..."). Figure 4.4 shows a selection of votes from the same game.

	creation_time	contents
0	2018-11-02 21:00:33.658168	Sarah Bryant: hi erybody
1	2018-11-02 21:00:39.856949	Julie Monroe: I bet it's Mandy. Mandy is an evil name
2	2018-11-02 21:00:40.196923	Mandy Smith: Hello
3	2018-11-02 21:00:48.892878	Mandy Smith: C'mon guy
4	2018-11-02 21:00:51.380136	Mandy Smith: I'm nice

Figure 4.3: Example messages (utterances) in a game. *creation_time* is the time at which the message was sent. The *contents* consists of the name of the sender, as well as the message, separated by a colon and space.

	creation_time	contents
0	2018-11-02 20:59:36.254353	Mandy Smith: Kevin Freeman
1	2018-11-02 21:01:09.581484	Sarah Bryant: Mandy Smith
2	2018-11-02 21:01:14.190603	Julie Monroe: Mandy Smith
3	2018-11-02 21:02:10.141132	Mandy Smith: Julie Monroe

Figure 4.4: Example of votes cast for eliminating players. *creation_time* is the time at which the vote was cast. *contents* consists of the name of the voter, as well as the player whom they wish to eliminate, separated by a colon and space.

4.4 Experiments

We begin by analyzing participants' empirical behavior as they play mafioso or bystander roles. First, we confirm that the mafia behave in line with their incentives, to validate that our experimental setup works. Next, we examine other emergent player strategies and behavior. Finally, we develop a classifier to differentiate between mafia and bystanders by analyzing their deceptive and deception-detecting behavior.

Previous studies have focused on the differences in message length between mafia and bystanders, suggesting that mafia tend to speak less than bystanders. However, mafia are also incentivized to influence group behavior through their speech. We thus hypothesized that differences in message length would be reflective of the balance between the mafia's desire to control group dynamics and their desire to not get caught in doing so. Prior work has also demonstrated that classifiers are able detect and use behavioral differences between mafia and bystanders. We therefore hypothesized that we should be able to build a classifier that was able to differentiate between bystanders and mafia members, and examine those measures in our results.

Validation of Game Paradigm

The first result we would expect in the Mafia game is that mafia members would try to eliminate more bystanders during the daytime phase than bystanders do, since the mafia are both incentivized to eliminate bystanders and know who the bystanders are. We test this hypothesis by examining the proportion of mafia member votes that are cast to eliminate a bystander, compared to the proportion of bystander votes cast for fellow bystanders. We observe that mafia do preferentially vote to eliminate bystanders compared to bystanders (Mafia $\mathrm{M} \pm \mathrm{SE} = 0.95 \pm 0.02$, Bystander $\mathrm{M} \pm \mathrm{SE} = 0.72 \pm 0.03$, $t(249.1) = 8.1$, $p < 0.0001$), validating that our game setup was working.

We would also expect that that mafia members would be more likely to vote in general during the daytime compared to bystanders. Mafia members have additional information on participants' roles, so are incentivized to use that information by voting to eliminate players. Even if Mafia members did not know who was a bystander, it would probabilistically make sense for them to vote more (since there are comparatively more bystanders than mafia, so any elimination is more likely to be a bystander) and Mafia members additionally have more practice with the voting system since they use it in the nighttime phase. To validate this hypothesis, we compare the averages for both groups of the proportion of daytime rounds in which a player casts a vote. As expected, we observe that mafia members vote significantly more during the daytime than bystanders do (Mafia $\mathrm{M} \pm \mathrm{SE} = 0.84 \pm 0.04$, Bystander $\mathrm{M} \pm \mathrm{SE} = 0.61 \pm 0.03$, $t(179.3) = 5.6$, $p < 0.0001$).

Having established that mafia members were engaging with the game paradigm as we would expect, we now examine what unexpected player strategies emerge.

Emergent Player Strategies

In a Mafia game, the mafia are outnumbered, but in return they have the advantage of information. The mafia know who the bystanders are, but the bystanders do not know who the mafia are, so it makes sense as a mafia member to talk a lot during the daytime to woo the majority opinion to their side. In this vein, we compare the average proportion of daytime rounds in which a player sends a message for each contingent. We find that mafia members are significantly more likely to send a daytime text than are bystanders (Mafia $\mathrm{M} \pm \mathrm{SE} = 0.84 \pm 0.04$, Bystander $\mathrm{M} \pm \mathrm{SE} = 0.60 \pm 0.03$, $t(187.9) = 6.2$, $p < 0.0001$), presumably due to their trying to influence the results by convincing bystanders of their opinions.

However, speaking more can pose a threat to players. As a bystander, there is little you can do to figure out who is a mafioso and who is a bystander, but watching the people who try to control the game is one of them. Sticking out can make you a target, as both mafia and bystanders seem to implicitly know: comparing the average proportion of votes that a player casts with the majority for both groups, both mafia and bystanders tend to conform in their voting behavior (Mafia $\mathrm{M} \pm \mathrm{SE} = 0.64 \pm 0.05$, Bystander $\mathrm{M} \pm \mathrm{SE} = 0.59 \pm 0.03$, $t(165.2) = 1.2$, $p = 0.22$). It is especially important to avoid being the first person targeted: if a player becomes the first person someone votes to remove during the daytime phase, there is a piling-on effect in which remaining players gang up on them, making it highly likely that that player will be eliminated (victim $\mathrm{rate} \pm \mathrm{SE} = 0.58 \pm 0.07$, expected $\mathrm{rate} = 0.12$, $t(48) = 10.5$, $p < 0.0001$)[2].

We hypothesized above that talkativeness might lead to elimination, but is this borne out in the data? We find that verbose talkers are indeed the players who get eliminated (victim $\mathrm{rate} \pm \mathrm{SE} = 0.43 \pm 0.07$, expected $\mathrm{rate} = 0.18$, $t(48) = 5.3$, $p < 0.0001$)[3]. Analogously, players who send the least texts during the daytime are eliminated at lower rates than expected (victim $\mathrm{rate} \pm \mathrm{SE} = 0.17 \pm 0.05$, expected $\mathrm{rate} = 0.38$, $t(48) = -5.3$, $p < 0.0001$).

Our findings suggest an interesting phenomenon with regards to incentives and deception. Since mafia members have more information about the game state than bystanders, in order to achieve their goals, they are incentivized to influence others' decisions by speaking more. However, bystanders are aware of this differential in knowledge and thus view such speakers as suspicious, hence requiring a trade-off between mafia members' ability to blend in by

[2]In this one-way t-test, the victim rate is the proportion of rounds in which the first proposed player is eliminated, averaged across all games. This victim rate is compared to the expected value. The expected value is calculated for each game by taking the number of first votes cast (equivalent to the number of rounds, e.g. 5), and dividing by the summed number of remaining (not-eliminated) players across each round (e.g. if 12 players start in the game, 11+9+7+5+3). The final expected value is the average of the game-specific expected values.

[3]In this one-way t-test, the victim rate is the proportion of rounds in which a player who sent the most number of texts is eliminated, averaged across all games. The expected rate is calculated as follows. First, we calculate the number of players who sent the highest number of texts each round (this should be one player per round, unless multiple players all tied), summed across rounds and averaged across all games. This value is divided by the summed number of remaining (not-eliminated) players across each round, averaged across all games.

speaking less, but to affect bystanders' actions by speaking more. It is important to note that mafia performed much better than bystanders, winning 31/49 (63%) of the games. This suggests that they are able to effectively balance between these competing pressures.

Classification of Player Role

Participants behave differently when they are playing a mafioso role compared to a bystander. However, was this difference in behavior large enough to build an automated classier that could determine who is a deceptive agent and who is not? By categorizing participants' texts and votes into basic features, we developed a classifier which could distinguish between mafia and bystanders.

We first featurized our data using the following: how often a participant cast a daytime vote, how often they cast a vote with the majority, how often they sent a daytime text, how often they sent a text before voting, the average time for them to cast a vote, the average time for them to send a text, the average number of texts they sent per round, and the average length in characters of a participant's texts.

Next, we randomly selected the 37 players of five games for our test set and used the 311 players of the remaining 44 games for training. We then over-sampled the mafia training data so that it was equal in number to that of the bystanders. Finally, we trained a random forest classifier on the resulting training set.

The classifier achieved an overall accuracy of 70.3%. In particular, we attained 71.4% accuracy for mafia classification. For comparison, bystanders correctly identified mafia in 30.1% of cases. Note that our classifier used information from all rounds of the game, whereas the bystanders could only use information up until the current round. However, even when given data corresponding to just the first daytime round, our model achieved 42.9% accuracy for the mafia, with 65.5% accuracy for the bystanders and 61.1% accuracy overall. In contrast, only 25.3% of bystander votes were cast for mafia members in the first round.

4.5 Discussion

Our analysis confirms that there are differences in behavior between honest and deceptive actors while playing the game of Mafia, even the limited text-based version we use here. These differences were salient to a classifier that could identify player roles with 70.3% accuracy. Our results are in line with previous work on training classifiers to identify deceptive roles. Comparing the performance of our classification model to humans, it appears that our model is able to contend with the mafia members' desire to speak more and influence bystanders while avoiding elimination. For example, when given information about participants' behavior through the second daytime round, humans tended to vote for the player who spoke the most, leading them astray in cases where said participant was a bystander. However, our model, while also assigning high probability of deception to such players, is

able to maintain higher probability on actual mafia members, suggesting that taking a more holistic approach to participant behavior can allow for correctly identifying deceptive actors in such situations.

We also found that mafia were significantly more likely than bystanders to vote for bystanders as opposed to fellow mafia members, as well as more likely to cast votes and send texts during the daytime. Previous work suggests that features such as linguistic diversity, the use of certain lexical items (eg. "but"), and the amount of language used can help to discern deceivers. Specifically, de Ruiter and Kachergis 2018 showed that mafia members sent fewer messages than bystanders for the Mafiascum dataset. Zhou and Sung 2008 also suggested the same pattern for a Chinese Mafia game. However, these results are in contrast to what we observed, wherein the mafia members were in fact more likely to send messages than bystanders.

While previous papers investigated differences in mafia versus bystander behavior, they did not analyze strategies used by participants to eliminate suspicious actors. Our results show that players who send the most texts in a round, as well as players who are first voted on to be eliminated in a round, are likely to be those who are ultimately selected to be eliminated. In contrast, those who are least talkative appear to be less likely to be eliminated, suggesting that this does not appear to be a trait of those who "stand out." It thus appears that bystanders do not randomly vote for who to eliminate, which is suggested as a strategy by Braverman et al. 2008 that would allow bystanders and mafia to win at equal rates. Instead, we observed that mafia win 1.7 times as many games as bystanders do.

One of our interesting findings was that, though players were more likely to eliminate those who spoke more, mafia were prone to speaking more anyway. This behavior may be explained by the game incentives which mirror those of the real world: if someone has additional information, they should leverage that knowledge to get people on their side while not raising undue suspicion. Since this behavior does not appear to result in mafia members losing the game more frequently than bystanders, this suggests a more complex relationship between strategies for deception and deception-detection. Future work should delve into this compensatory mechanism wherein the amount of information different parties have is varied in a controlled experiment.

4.6 Conclusion

In an environment where one party has hidden information, that party should talk more to convince others to do as they wish. However, the other people should be more likely to punish those who talk more as a compensatory mechanism for this incentive structure. This is exactly what we found in examining participants' behavior in the game of Mafia: mafia spoke more than bystanders, but bystanders and mafia alike punished and tended to eliminate participants who spoke more. This finding illustrates a fascinating interplay of deception and detection that emerged naturally in gameplay. Moreover, these differences in how participants behaved in different roles are not just descriptive: we developed a classifier

to determine whether, based information about texting and voting behavior, our computational model could automatically distinguish mafia and bystanders. Indeed, the model could predict who was a mafia member and who was a bystander with 70.3% accuracy using a simple set of 8 features. These results illustrate some of the mechanisms of the implicit back-and-forth nature of deception and detection of deception in human interaction, and also test the extent of this knowledge by determining that an automatic detection system can indeed determine the roles that a participant is playing. In addition to providing insight on how such entities use deception to avoid detection, this may have implications for online communication paradigms, providing possible avenues for verification of unknown entities within a network in the real world.

Chapter 5

Putting the *Con* in Context: Identifying Deceptive Actors in the Game of Mafia

In the last chapter, we introduced a framework to collect a dataset of Mafia game records, using this framework to investigate strategies used by players of the game in the roles of deceiving and detecting deception in others. In this chapter, we further analyze the effect of speaker role on language use through the game of Mafia, demonstrating that there are differences in the language produced by players with different roles. We confirm that classification models are able to rank deceptive players as more suspicious than honest ones based only on their use of language. Furthermore, we show that training models on two auxiliary tasks outperforms a standard BERT-based text classification approach. We also present methods for using our trained models to identify features that distinguish between player roles, which could be used to assist players during the Mafia game.

5.1 Introduction

This work explores language used for deception: a type of speaker context that is particularly challenging to model because it is intentionally hidden by the speaker. To do so, we collect and release a set of records for the game of Mafia[1], in which each player is assigned either an honest or a deceptive role. Then, we develop models that distinguish players' roles based only on the text of the players' dialog. We describe two auxiliary tasks that improve classification accuracy over a BERT-based text classifier.

The novel contributions of this work include:

1. A methodology for collecting records of online Mafia games and a dataset collected from 460 human subjects,

[1]Dataset is available at `https://paperswithcode.com/dataset/the-mafia-dataset`.

	Mafia	Bystander	Total
Total number of players	87	334	421
Average number of players per game	1.98	7.59	9.57
Total number of day utterances	443	1382	1825
Average number of day utterances per player	5.1	4.1	4.3
Total number of players without day utterances	7	71	78
Percent of players without day utterances	9.0%	91.0%	100%

Table 5.1: Mafia dataset statistics. **Mafia** and **Bystander** denote the mafia and bystander classes, respectively, while **Total** denotes the total number for both groups. Since bystanders are unable to talk during the nighttime, we only show statistics for daytime utterances (*day* is short for *daytime*). The last row shows the distribution of roles among the players with no utterances throughout the game. Note that nearly all of the no-utterance players are bystanders.

2. Three classification models that can distinguish between honest and deceptive players,

3. An approach for identifying features of the game dialog text that can be used to help identify deceptive players during the game.

The task of identifying deception in dialog is far from solved. Our classification methods, while not accurate enough to reliably identify deceptive players in a game, do show that the text of a dialog in the setting we study does contain information about the roles of the participants, even when those participants are motivated to hide those characteristics by deceiving the listener. Although the models and results described in this work only apply to a particular game setting rather than dialog in general, the approaches we describe are general in character and therefore may inform future work on determining speaker roles from the contents of dialog.

5.2 Dataset

We use the same dataset as described in Chapter 4, discarding five of the Mafia games for quality, which left forty-four games for the final analysis. Dataset statistics appear in Table 5.1.

Upon further inspection of the data, we can observe several strategies used by mafia members to deceive bystanders:

1. Mafia members may suggest that there is not enough information to decide on who to eliminate, despite their knowledge of everyone's roles (eg. "Should we wait to eliminate someone?" / "It's a little early to tell." / "It's a shot in the dark."),

2. Mafia members may raise suspicion about another player, despite knowing that said player is a bystander (eg. hmm ok analyzing this conversation....I think bianca was a little to flippant in how she was like "sucks to be andrew" haha / I'm going to vote bianca. she's so casual with life and death),

3. Mafia members may invent a false motive and assign that motive to another player, despite knowing that the player is a bystander (eg. It might be Jonathan Kim... killing off Erin who accused him "yesterday").

5.3 Approach

Given our mafia dataset, there are several tasks that one might address, for example, predicting participants' daytime voting behavior or generating mafia members' nighttime dialog. As our aim is to identify deceptive actors, however, we focus on predicting participants' roles, i.e. bystander or mafioso. Due to the asymmetry in the knowledge available to each group and the goals which incentivize bystanders to increase true belief and mafia members to reduce it, the bystanders are said to take on an honest role in the game, whereas the mafia members take on a deceptive role. To focus on the relationship between language and deception, we ignore voting behavior and consider just the daytime dialog in the game, as only the mafia members were able to converse during the nighttime. As shown in Table 5.1, since most of the players with no utterances are bystanders, we only consider players who make at least one utterance throughout the game.

To investigate whether linguistic information can be used to identify players' roles, we train and evaluate classifiers that predict the role of a particular player. Since we have a small dataset, we chose to fine-tune pre-trained Transformer models rather than train them from scratch (Vaswani et al., 2017). To predict the role for a player p, we construct an input representation $r(C, p)$ of the full game dialog C that encodes the player of interest p. We develop three approaches which differ in both the dialog representation function r and the modeling approach.

Standard Classification

Our baseline approach uses a standard BERT-based text classifier (Devlin et al., 2018). To classify player p via the full record of the game C, let boolean variable M_p be true if p is a mafioso. Let T_p be the concatenation[2] of utterances made by p. We train BERT parameters θ_M to predict $P(M_p|T_p; \theta_M)$.

This approach, which provides as input to the classifier only the utterances of the player to be classified, outperformed an alternative representation $r(C, p)$ that included the entire record of all utterances by all players.

[2]Utterances are concatenated with an end-of-sentence delimiter after each utterance.

Auxiliary Tasks

Limiting the input representation r to contain only the speech of the player p being classified is not ideal; correctly interpreting a dialog requires considering all other players' statements as well. We introduce two auxiliary tasks that involve the entire game dialog C:

1. Given all of the prior utterances, is a bystander or a mafia member more likely to have produced the current utterance? (*Utterance Classification*)

2. Given all of the prior utterances, what current utterance would a player produce, given that they are a bystander or a mafia member? (*Utterance Generation*)

We develop a BERT-based classification model for task 1 and fine-tune the GPT-2 language model for task 2 (Radford et al., 2019). Then, we use each of these auxiliary models to classify the role of a particular player p in the game.

Utterance Classification

To classify player p using the auxiliary task of utterance classification, let boolean variable S_i be true if utterance C_i was made by a mafioso (rather than a bystander). Let C be the full record of utterances in the game and $C_{\leq i}$ be the concatenation of all utterances $C_1 \dots C_i$. We train BERT parameters θ_S to predict $P(S_i|C_{\leq i}; \theta_S)$. Finally, let I_p be the set of indices of utterances by player p. M relates to S in that if M_p is true, then S_i is true for all $i \in I_p$. We thus calculate

$$P(M_p|C; \theta_S) \propto \frac{\sum_{i \in I_p} P(S_i|C_{\leq i}; \theta_S)}{N},$$

where $N = |I_p|$.

Utterance Generation

To classify player p using the auxiliary task of utterance generation, we fine-tune GPT-2 to generate utterance C_i conditioned on prior utterances $C_{<i}$ and the role S_i of the speaker that produced C_i. From Bayes' rule, we have $P(M_p|C) \propto P(M_p)P(C|M_p)$. To estimate $P(C|M_p)$, let C_p include all C_i for $i \in I_p$. We make the simplifying assumption that $P(C|M_p) \propto P(C_p|M_p)$, which assumes that the utterances made by players other than p are independent of the role of player p. Then, if M_p is true, S_i is true for all $i \in I_p$, and so,

$$P(C_p|M_p; \theta_C) = \prod_{i \in I_p} P(C_i|C_{<i}, S_i; \theta_C).$$

Using the full dialog C, the final probability of player p being mafioso is calculated as follows:

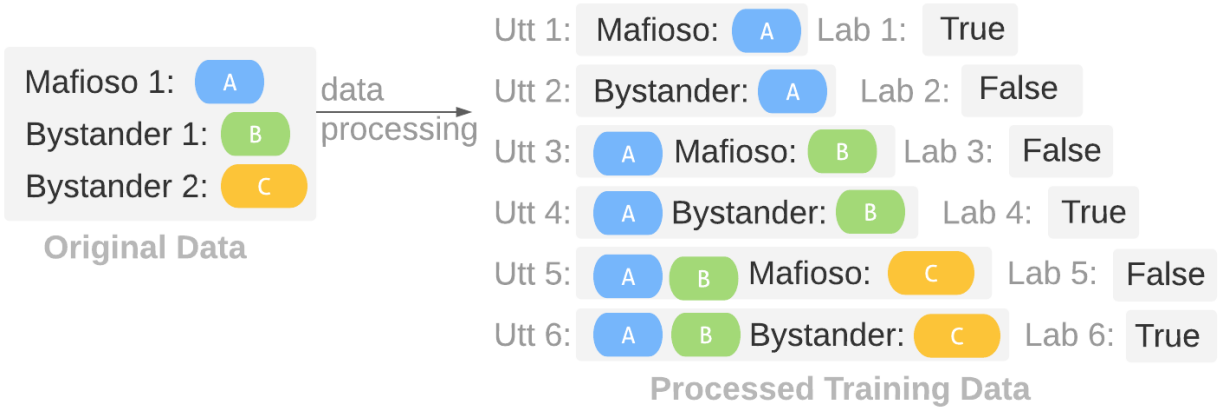


Figure 5.1: Data processing for fine-tuning BERT model. The original data is shown on the left-hand side, while the right-hand side shows the processed data containing two versions of each utterance, one assuming that the target player is a mafioso and one assuming that they are a bystander, with the prior conversation context preceding each and labels corresponding to whether the assumed role matches the actual role of the player.

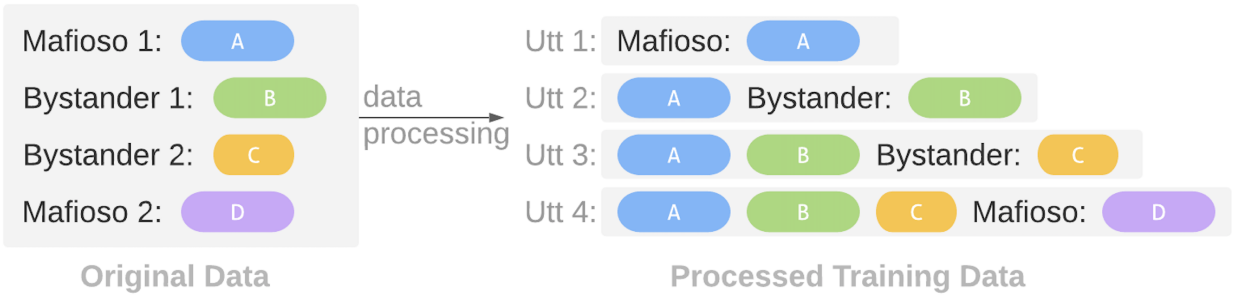


Figure 5.2: Data processing for fine-tuning GPT-2 model. The original data is shown on the left-hand side, while the right-hand side shows the processed data containing a version of the corresponding utterance with the prior conversation context preceding.

$$P(M_p|C) = \frac{P(M_p)P(C_p|M_p;\theta_C)}{\sum_{R\in\{M,\neg M\}} P(R_p)P(C_p|R_p;\theta_C)} \tag{5.1}$$

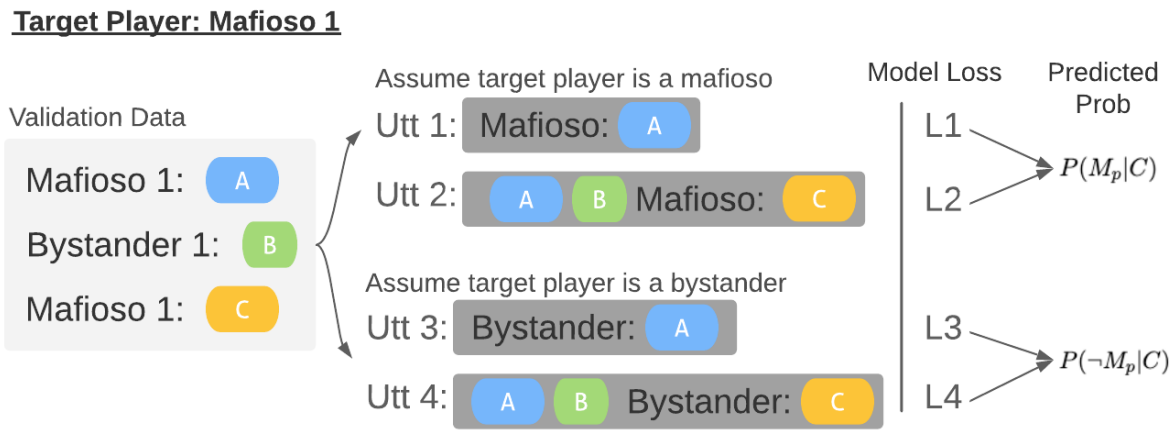


Figure 5.3: Prediction pipeline for our fine-tuned GPT-2 model. Similar to the pipeline used to produce the training utterances, for prediction, there are now two versions of each, one assuming that the target player is a mafioso and one assuming that they are a bystander. The losses for each utterance of the target player are summed together in order to calculate the mafia and bystander probabilities as described in Equation 5.1.

Data Processing

To train models for utterance classification (using BERT) and utterance generation (using GPT-2), we perform data processing procedures on the games' original dataset to create input representations $r(C, p)$ for each player p and obtain our training datasets as shown in Figures 5.1 and 5.2. The left side of each figure shows a snippet of a game's data, where "Mafioso" and "Bystander" denote the true roles of the players. The utterances to the right of each figure are training examples used for fine-tuning the BERT and GPT-2 models. Structuring the data in this way provides both the prior context of utterances and the current utterance that happened within this context. This not only gives us the information needed for the auxiliary tasks, but also provides us with more training examples, as we only have 44 games and only 421 players in total, with only 2162 total utterances. Moreover, this mimics the real game scenario from the bystander view in that they can only confirm their own role, but no one else's, which is the appropriate setting for us in which to detect deception.

Figure 5.3 shows the pipeline for using the GPT-2 model to predict players' roles. Let us assume that the target player for whom we want to predict their role is Mafioso 1. From the original game log on the left, we first perform the data processing scheme from Figure 5.2 twice, assuming that the target player is a mafioso (top of Figure 5.3) and a bystander (bottom of Figure 5.3). Using our trained GPT-2 model, we then obtain a loss for each

	Avg Rank	Avg Rank/Game	Accuracy	Maf F1	Bys F1
Random	19.0	3.4	0.62	0.26	0.74
Std Class	17.9	3.0	0.69	0.4	0.79
Utt Class	14.5	**1.8**	**0.74**	**0.50**	**0.83**
Utt Gen	**11.4**	2.0	**0.74**	**0.50**	**0.83**

Table 5.2: Experiment results on the validation set for random baseline (**Random**), standard classification (**Std Class**), utterance classification (**Utt Class**), and utterance generation (**Utt Gen**) approaches. Methods that use auxiliary tasks (**Utt Class** and **Utt Gen**) outperform other methods in terms of average ranking overall and per game while also maintaining higher accuracy and F1-score for each class.

utterance denoted by L1 through L4. Summing all the losses for each role, as they denote log probabilities, we calculate $P(M_p|C)$ and $P(\neg M_p|C)$ via Equation 5.1. The target player's role as predicted by the model is finally given by comparing the two probabilities. A similar process is used to calculate $P(M_p|C)$ and $P(\neg M_p|C)$ for the utterance classification BERT model.

5.4 Experiments

We train three fine-tuned models on the corpus of Mafia game records and compare their performance to a random baseline. The specifications for the baseline and models can be found below, and the results are shown in Table 5.2.

Random Baseline

This random classifier classifies each player as a mafioso or a bystander with probabilities equal to the prior distribution of each class, estimated as the ratio of roles across all training games. This serves as a baseline to be compared to for all other methods. In the game setting, this mimics a bystander player with only public information of how many mafia and bystanders are in the game.

Standard Classification

We initialize the model by loading a pre-trained BERT Base model (12 layers, 768 hidden dimension size, 12 attention heads). We train with a maximum sequence length of 256, which is sufficient for our post-processed dataset, setting the batch size to 16, the learning rate to 1e-5, and the maximum number of epochs to 25.

Utterance Classification

We initialize the model by loading a pre-trained BERT Base model (12 layers, 768 hidden dimension size, 12 attention heads). We train with a maximum sequence length of 512, which is sufficient for our post-processed dataset, setting the batch size to 5, the learning rate to 5e-5, and the maximum number of epochs to 25.

Utterance Generation

We initialize the model by loading a pre-trained 12-layer GPT-2 model with an embedding size of 768. For the dataset, we set the maximum length of each sentence to be 512, which is sufficient for our dataset after post-processing. During training, we set the batch size to be 5 and the learning rate to be 1e-5. We train the model for a maximum of 100 epochs.

Metrics

These approaches each estimate a probability $P(M_p|C)$ that a player p is a mafioso given the full record of game texts C. In Mafia, bystanders do not declare who is and is not a mafioso, but instead vote each day to eliminate one of the players. Because the act of voting involves choosing one player among them all, a natural metric for evaluating the usefulness of a model is to order all players p from greatest to least $P(M_p|C)$, their probability of being a mafioso under the model, and then to compute the average rank of the true mafia members. Therefore, the first metric in Table 5.2 is the average ranking of all mafia members when each player is ranked by $P(M_p|C)$ across the entire validation set composed of 5 games. It is also natural to consider player ranking within a single game, so we calculate the average ranking of mafia members within each game as a second metric. Smaller average ranking for mafia members means that the model is able to assign mafia players a high $P(M_p|C)$ relative to bystanders, which is desired.

In addition, we evaluate the accuracy of the classifiers and the F1-score for each class. To calculate these metrics, we first assign the mafioso label to the top k players with the highest $P(M_p|C)$ and the rest of the players with the bystander label, where k is the known number of mafia among all validation games ($k = 10$ in our case). Aside from the ranking metrics, these give further information of the models' quality after utilizing available game information.

Results and Analysis

We trained all models on 39 training games and evaluated on the remaining 5 validation games. The evaluation results are shown in Table 5.2. We have a total of 49 players in the validation games, but only considered the 39 players who had spoken at least one utterance throughout the game when calculating the metrics. Players with no utterances are almost exclusively bystanders and are therefore easy to classify without considering language.

First, we see that it is possible to achieve an average rank that is smaller than the random baseline, which demonstrates that there is information in the dialog about the roles of players, despite the fact that mafia members seek to hide their role while conversing. However, standard classification is comparable to random. Next, we observe that both models using auxiliary tasks outperform the standard classifier in rank-based metrics, which demonstrates that the auxiliary tasks provide useful inductive bias for the mafia classification task. Additionally, the accuracy is similar for all approaches, including random classification, which indicates that there is not enough information in the text of a Mafia game for these models to determine players' roles reliably. If the goal of the game were to guess the role of each player individually, then always guessing bystander (i.e. the majority class) would be the best strategy. However, since the goal for the bystanders is to vote to eliminate a mafia member each round, the utterance classification and utterance generation approaches, which achieve the lowest average mafia ranking per game and overall, respectively, are the most favorable.

Note that the precision for the mafia is much lower than that of the bystanders for all models. This is due to the usual lack of information available to predict that any player is a mafioso, which makes finding the mafia a much harder task than finding bystanders.

5.5 Discussion

The decoding ability of the GPT-2 model provides us a more straightforward way to understand what the model has learned. Given a prompt sentence, we can use our fine-tuned GPT-2 model to generate what a mafioso and a bystander would say. A few examples are shown in Table 5.3. From these examples, we inspect the following features that the model might be capturing to distinguish between mafia and bystanders: Feature 1: Referring to other players. Feature 2: Expressing confusion. Feature 3: Referring to others for elimination purposes. Feature 4: Asking for suggestions on who to eliminate.

To confirm that our fine-tuned GPT-2 model captures some of the above features, we hand-label these features on 5 training games and 1 validation game, obtain each player's feature vector, and see whether there exists a correlation between the model's predicted $P(M_p|C)$ for validation players and the similarity of their feature vectors compared to the training set mafioso and bystander players. These feature vectors are shown in Table 5.4, where each entry denotes the average number of features per player of each role. As an example, for the first column, each mafioso player says 2 utterances having Feature 1 throughout the game on average, while each bystander player says 1.06 utterances having Feature 1 on average. We define the first row as a vector v_1 and the second row as v_2 for future references.

Table 5.5 shows the hand-labeled feature vectors for all 10 players in a validation game (first 4 columns, F1 to F4) ranked by the model's predicted $P(M_p|C)$. We define a metric function $D(u) = \|u - v_1\|^2 - \|u - v_2\|^2$ for a validation player's feature vector u. The smaller $D(u)$ is, the closer u is to v_1 than v_2, and hence the more mafia-like they are with respect to players in the training games. We can see that for players of higher rank, their $D(u)$

Prompt	Generated Utterance
lets kill P1.	**M:** sorry P1 :(**B:** hello all
who thinks P3 is Mafia?	**M:** No i'm a bystander **B:** No idea
That sounds suspicious...	**M:** P6 is mafia **B:** Why yall want to eliminate me?
hi team. Hello!. Hi.	**M:** Who is the mob person? **B:** hello

Table 5.3: Utterances generated by our GPT-2 model given different prompts. **M** and **B** are shorthand for Mafioso and Bystander respectively, and P1, P3, and P6 denote the names of other players in the game.

	Feat 1	Feat 2	Feat 3	Feat 4
Mafioso	2.00	0.00	1.30	0.40
Bystander	1.06	0.27	0.65	0.10

Table 5.4: The average count per role for each of four hand-labeled features (number of references to other players, level of confusion, number of references to other players for elimination, and number of requests for who to eliminate) as identified by our GPT-2 model on 5 training games.

are negative with larger magnitudes. Referring to the true labels in the rightmost column (M for Mafioso and B for Bystander), the first row also explains how our model can fail to predict the true role of some players: even though this player is a bystander, they act more like the mafia than other bystanders according to these hand-labeled features because they are regularly referencing and accusing other players.

5.6 Conclusion

The widespread use of the internet by general users has allowed for a myriad of advantages, including the ability to communicate with people who are physically distant from oneself. However, this has rendered the same users as more vulnerable to deceptive parties, who are no longer limited to those who are in close proximity to them. In order to explore possible protections against such deception, we investigated how extrinsic speaker attributes such as conversational role affect language in the game of Mafia. By leveraging this environment for which roles are explicitly labelled, as well as incorporating auxiliary tasks that model language in context, we are able to make progress toward the task of deception detection, an essential method to protect users in a world that is becoming progressively more online.

	F1	F2	F3	F4	D(u)	Pred	Truth
P0	4	0	2	0	-5.9	0.98	B
P1	2	0	2	0	-2.1	0.93	M
P2	5	0	5	0	-11.7	0.78	M
P3	2	0	2	0	-2.1	0.63	B
P4	4	2	1	1	-4.1	0.47	B
P5	3	0	2	0	-4.0	0.43	B
P6	0	0	0	0	4.2	0.42	B
P7	1	0	1	0	1.0	0.40	B
P8	0	0	0	0	4.2	0.00	B
P9	0	0	0	0	4.2	0.00	B

Table 5.5: Features for each player (P0 to P9) in a validation game. For each row, F1 to F4 give the feature vector u for the respective player. $D(u)$ gives the similarity of u compared to the training feature vectors v_1 and v_2. Players are sorted by $Pred$, the probability $P(M_p|C)$ given by our GPT-2 model, and $Truth$ gives the true label for each player (M for Mafioso, B for Bystander). Since P8 and P9 have no utterances throughout the game, as per our heuristic, they are predicted to be bystanders with $P(M_p|C) = 0$.

Chapter 6

Conclusion

In this thesis, we have investigated methods for leveraging speaker context for NLP tasks including text classification and text generation. We show that models may be trained to identify author gender on the Reddit platform, as well as to obfuscate the gender of such authors from being determined by their comments. Further, we demonstrate that we can train classifiers to distinguish between honest and deceptive actors in the game of Mafia, using context-aware tasks to improve model performance.

For our work on gender obfuscation, we rely on an imperfect gender classification model. Future work should investigate how this model's accuracy affects gender obfuscation performance. We also rely on automatic techniques to evaluate our gender obfuscation approach. As our focus was to protect authors' genders from being discerned by classifiers similar to the one we developed, such automatically calculated metrics suffice for the scope of our work. However, in the future, a human evaluation could be implemented to determine how well such gender obfuscation approaches would perform against human adversaries.

As of now, these models are targeted paraphrasing methods that focus on small phrases to reduce the change to the original sentence. They can be used iteratively if a user wants to accumulate multiple changes over a span of text, but this still leaves issues of consistency that can arise due to such targeted methods. For example, lacking other obvious signals, the genders of other people in a sentence may be changed to move the gender prediction. However, because these are targeted methods, they do not account for references to the same entity, so we may get examples where "my girlfriend" is changed to "my boyfriend", but then is later on referred to as "she". Future work should investigate methods for correcting inconsistencies that can be introduced through these gender obfuscation edits, as well as methods for maintaining consistency across multiple gender obfuscation edits.

Furthermore, methods of this nature are often limited by what signals they can pick up in the data. These signals may not always align with what we consider to be correct interpretations of the world. For such gender obfuscation methods, the models try to identify the most salient information to mark someone's gender, but these may not accurately reflect our current views. For example, most of the relationship terms for Reddit accounts that specify their gender fall into a heteronormative worldview. That is, male and female accounts

tend to refer to their spouses as wives and husbands, respectively. This means that for a piece of text where the author's partner is mentioned, one of the easiest ways for the model to change the perceived gender is simply to adjust the term of the partner to follow heteronormative standards, whereas to preserve meaning we would likely prefer defaulting to a gender-neutral term like 'partner' or 'significant other'. As the aim of this work is to produce gender-neutral text, there is also the potential risk that users may assume this approach would be able to generate text that is consistent with the language of non-binary individuals. Though our methods are general in nature, they have not been evaluated against this kind of language in particular, so future work should investigate whether these models can be adapted to such contexts. Finally, while obfuscating an individual's gender can help preserve privacy and make them feel safer, it is a more short-term solution to general issues of representation and acceptance. In order to help mitigate against these ethical issues in automatically neutralizing gendered language, the decision on how to employ such a gender obfuscation approach should ultimately be delegated to the user.

As for our Mafia analysis, we observed that after one participant proposed a person to eliminate, other participants tended to follow this suggestion. We also observed cases in which multiple suggestions were made for who to eliminate. Future work should further investigate how group dynamics factor into these elimination suggestions, for example examining situations in which both a mafia member and a bystander are proposed as possible victims to provide insight into how well participants can differentiate between these roles. Such work can also ask how people, particularly mafia members, successfully defend themselves from accusations to explore not only models for detecting deception, but also models for deceiving others. We were able to train models to help differentiate players with different roles in the game of Mafia based only on their language use, as well as to identify features that may distinguish between these roles. We also noticed that the mafia were twice as likely to win the Mafia game than were the bystanders. These findings lead us to believe that the bystanders may benefit from being provided hints based on our model's predictions and identified features, or that our models may be trained to participate in the game themselves.

However, there are several ethical considerations in regards to using these methods. First, as our model is trained on this particular version of mafia, the specific models trained would not apply to other cases of deceptive language use. Applying these models to out-of-domain data, or even adapting this general approach to new settings, may yield unexpected results. Our experimental results only establish the effectiveness of our approach on the game of Mafia. Future work must evaluate these approaches on other deception detection tasks before they can be safely deployed in real-world scenarios. Next, information that may aid bystanders in detecting deception may also aid mafia members in being deceptive. Though mafia members may attempt to use it for this purpose, because our model is trained to increase true belief, which is directly in line with the bystander goal to identify the truth and against the mafia goal to obscure it, our approach is inherently more useful to bystanders. However, since the models we evaluate are far from perfectly accurate, there is a risk that users using these models for hints would rely too much on their output and thereby be misled. More work should be done to increase the model's performance in order to mitigate

this risk.

Language as a communication channel allows users to express various forms of surrounding context in addition to the content that they wish to convey. By introducing methods for natural language processing systems to take greater advantage of this context in terms of both intrinsic speaker attributes such as gender and extrinsic speaker attributes such as conversational role, we hope that this contributes to future avenues for these systems to provide users increased agency and security in their online interactions.

Bibliography

Mohamed Abouelenien, Veronica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. Deception detection using a multimodal approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 58–65, 2014.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*, 2016.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. User-centric gender rewriting. *arXiv preprint arXiv:2205.02211*, 2022.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839, 2020.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108, 2019.

Xiaoheng Bi and Tetsuro Tanaka. Human-side strategies in the Werewolf game against the stealth werewolf strategy. In *International Conference on Computers and Games*, pages 93–102. Springer, 2016.

Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv preprint arXiv:1707.00061*, 2017.

Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234, 2006.

Mark Braverman, Omid Etesami, and Elchanan Mossel. Mafia: A theoretical study of players and coalitions in a partial information environment. *The Annals of Applied Probability*, 18(3):825–846, 2008.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Judee K Burgoon, J Pete Blair, Tiantian Qin, and Jay F Nunamaker. Detecting deception through linguistic analysis. In *International Conference on Intelligence and Security Informatics*, pages 91–101. Springer, 2003.

Judith Butler. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre Journal*, 40(4):519–531, 1988.

Gokul Chittaranjan and Hayley Hung. Are you a werewolf? Detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? An analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Robin De Pril. User classification based on public Reddit data. *Ghent University.: Ghent University*, 2019.

Bob de Ruiter and George Kachergis. The Mafiascum Dataset: A large text corpus for deception detection. *arXiv preprint arXiv:1811.07851*, 2018.

Sergey Demyanov, James Bailey, Kotagiri Ramamohanarao, and Christopher Leckie. Detection of deception in the Mafia party game. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 335–342, 2015.

Douglas C Derrick, Thomas O Meservy, Jeffrey L Jenkins, Judee K Burgoon, and Jay F Nunamaker Jr. Detecting deceptive chat-based communication using typing behavior and message cues. *ACM Transactions on Management Information Systems (TMIS)*, 4(2): 1–21, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*, 2020.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. Refocusing on relevance: Personalization in NLG. *arXiv preprint arXiv:2109.05140*, 2021.

Chris Emmery, Enrique Manjavacas, and Grzegorz Chrupała. Style obfuscation by invariance. *arXiv preprint arXiv:1805.07143*, 2018.

Tommaso Fornaciari and Massimo Poesio. Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, 21(3):303–340, 2013.

Christie M Fuller, David P Biros, and Dursun Delen. An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications*, 38 (7):8392–8398, 2011.

Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

Samee Ibraheem, Vael Gates, John DeNero, and Tom Griffiths. Investigating the behavior of malicious actors through the game of mafia. In *CogSci*, 2020.

Samee Ibraheem, Gaoyue Zhou, and John DeNero. Putting the con in context: Identifying deceptive actors in the game of mafia. *arXiv preprint arXiv:2207.02253*, 2022.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Michael Kearns, Stephen Judd, Jinsong Tan, and Jennifer Wortman. Behavioral experiments on biased voting in networks. In *Proceedings of the National Academy of Sciences 106.5*, pages 1347–1352, 2009.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? End-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*, 2017.

Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song. Unsupervised paraphrasing by simulated annealing. *arXiv preprint arXiv:1909.03588*, 2019.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, 2017.

Piotr Migdał. A mathematical model of the Mafia game. *arXiv preprint arXiv:1009.1031*, 2010.

Rada Mihalcea, Verónica Pérez-Rosas, and Mihai Burzo. Automatic detection of deceit in verbal communication. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 131–134, 2013.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. JFLEG: A fluency corpus and benchmark for grammatical error correction. *arXiv preprint arXiv:1702.04066*, 2017.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45 (3):211–236, 2008.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. Linguistic harbingers of betrayal: A case study on an online strategy game. *arXiv preprint arXiv:1506.04744*, 2015.

Jinie Pak and Lina Zhou. A social network based analysis of deceptive communication in online chat. In *Workshop on E-Business*, pages 55–65. Springer, 2011.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. Debiasing embeddings for reduced gender bias in text classification. *arXiv preprint arXiv:1908.02810*, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sravana Reddy and Kevin Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, 2016.

Aurko Roy and David Grangier. Unsupervised paraphrasing without translation. *arXiv preprint arXiv:1905.12752*, 2019.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, 2016.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*, 2019.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*, 2021.

Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, 2017.

Rachael Tatman and Conner Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *INTERSPEECH*, pages 934–938, 2017.

Mike Thelwall and Emma Stuart. She's reddit: A source of statistically significant gendered interest information? *Information processing & management*, 56(4):1543–1558, 2019.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives. *arXiv preprint arXiv:2109.06105*, 2021.

Evgenii Vasilev. Inferring gender of Reddit users. 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Charles Welch, Chenxi Gu, Jonathan Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, 2022.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised domain adaptation for neural machine translation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 338–343. IEEE, 2018.

Lina Zhou and Yu-wei Sung. Cues to deception in online Chinese groups. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, pages 146–146. IEEE, 2008.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, 2019.

Appendix A

Mafia Instructions

Below is a transcript of the instructions that were provided to participants before playing the Mafia game in our experiments:

"In this experiment, you will play a version of the party game "Mafia". You are going to play the game of Mafia (also known as Werewolf) with other participants. You are either part of the mafia (a mafioso) or a bystander. The mafia will know who is in the mafia, but the bystanders will not. There will always initially be more bystanders than mafia. There will be one or more mafia members. The goal of the mafia is to eliminate the bystanders one by one until the mafia are equal in number to them. The goal of the bystanders is to correctly guess the identity of the mafia and eliminate them all before the mafia win. There are two phases to this game, nighttime and daytime; at the end of each, a participant is eliminated from the game:

1. In the **nighttime** phase, only the mafia can converse and decide who they want to eliminate. Specifically, if you are a mafioso, you will talk in a chatroom, then use a drop-down menu to select who you want to remove. Mafia will have 1 minute to do this. If there is more than one mafioso and the mafia disagree about who to eliminate, one of the mafia's choices will be selected randomly. If you are a bystander, you will wait out this time, as you are sleeping during the night.

2. Everyone is awake during the **daytime** phase. The participant who was eliminated during the night will be announced: if you were eliminated, you will be sent to the end of the game and compensated. The remaining participants will converse (for 2 minutes and 30 seconds) and decide who to eliminate, where the goal of the bystanders is to eliminate a member of the mafia, and the goal of the mafia is to eliminate a bystander. By the end of this time, everyone needs to select a name from the drop-down menu. (If there are multiple mafia, the mafia will be reminded of each others' names in separate text on this page.) The participant with the most votes will be eliminated, except in the case of a tie, in which a randomly-selected vote will be eliminated. The eliminated participant and their identity (bystander or mafia) will be announced, and that participant will be sent to the end of the game and compensated.

The game will continue, alternating between nighttime and daytime, until either all of the mafia are removed (*bystanders win!*) or there are equal numbers of mafia and bystanders (*mafia win!*)"