Chapter 1: See Other File

Chapter 2: See Other File

Chapter 3: See Other File

Chapter 4: See Other File

Chapter 5: See Other File

Chapter 6: See Other File

Chapter 7: Expert *Generation*

Experts can **generate** datasets of a quality unachievable by the **crowd**. This has a twofold benefit. First, the accuracy, rather than the size, of the data allows the dataset to withstand the test of time.[1] Second, tasks that require long-term commitment and complexity become possible. This result justifies the large investment of time, relationship-building, and money necessary to use experts.

We create a deception dataset using experts, as a contrast to the earlier crowd-sourced generated CANARD dataset (Chapter **??**), Participants—that are engaged in the task and are appropriately compensated—both generate and annotate data in the span of a game that usually lasts over a month. The **annotation** is more complicated than in our adaptation dataset (Chapter **??**) due to being real-time and user-specific. The resulting product is a gold standard of conversational NLP data in quality of language, diversity, and naturalness.[2] The conversations and annotations

[1]The Penn Treebank (Marcus et al., 1993), which used graduate students in linguistics and spanned three years in the early 1990s, remains a staple of Computational Linguistics curriculum today

[2]Denis Peskov, Benny Chang, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It Takes Two to Lie: One to Lie and One to Listen. In Proceedings of The Association for Computational Linguistics. Peskov was responsible for designing the task, gathering the participants, running the games, building half the models, part of the data analysis,

thereof would not be possible without experts familiar with the game.

7.1 Where Does One Find Long-Term Deception?

A functioning society is impossible without trust. In online text interactions, users are typically trusting (Shneiderman, 2000), but this trust can be betrayed through false identities on dating sites (Toma and Hancock, 2012), spearphishing attacks (Dhamija et al., 2006), sockpuppetry (Kumar et al., 2017) and, more broadly, disinformation campaigns (Kumar and Shah, 2018). Beyond such one-off antisocial acts directed at strangers, deception can also occur in sustained relationships, where it can be strategically combined with truthfulness to advance a long-term objective (Cornwell and Lundgren, 2001; Kaplar and Gordon, 2004).

We introduce a dataset to study the strategic use of deception in long-lasting relationships. To collect reliable ground truth in this complex scenario, we design an interface for players to naturally **generate** and **annotate** conversational data while playing a negotiation-based game called Diplomacy. These annotations are done in *real-time* as the players send and receive messages. While this game setup might not directly translate to real-world situations, it enables computational frameworks for studying deception in a complex social context while avoiding privacy issues.

After providing background on the game of Diplomacy and our intended deception annotations (Section 7.2), we discuss our study (Section 7.4). To probe the value of the resulting dataset, we develop lie prediction models (Section 7.5) and an-

the visualizations, and the paper writing.

alyze their results (Section 7.6). The role of the **expert** is paramount (Section 7.9).

7.2 Diplomacy

The Diplomacy board game places a player in the role of one of seven European powers on the eve of World War I. The goal is to conquer a simplified map of Europe by ordering armies in the field against rivals. Victory points determine the success of a player and allow them to build additional armies; the player who can gain and maintain the highest number of points wins.[3] The mechanics of the game are simple and deterministic: armies, represented as figures on a given territory, can only move to adjacent spots and the side with the most armies always wins in a disputed move. The game movements become publicly available to all players after the end of a turn.

Because the game is deterministic and everyone begins with an equal amount of armies, a player cannot win the game without forming alliances with other players—hence the name of the game: Diplomacy. Conquering neighboring territories depends on support from another player's armies. After an alliance has outlived its usefulness, a player often dramatically breaks it to take advantage of their erstwhile ally's vulnerability. Table 7.1 shows the end of one such relationship. As in real life, to succeed a betrayal must be a surprise to the victim. Thus, players pride themselves on being able to lie *and* detect lies. Our study uses their skill and pas-

[3]In the parlance of Diplomacy games, points are "supply centers" in specific territories (e.g., London). Having more supply centers allows a player to build more armies and win the game by capturing more than half of the 34 supply centers on the board.

sion to build a dataset of deception created by battle-hardened diplomats. Senders annotate whether each message they write is an ACTUAL LIE and recipients annotate whether each message received is a SUSPECTED LIE. Further details on the annotation process are in Section 7.4.1.

7.2.1 A game walk-through

Figure 7.1 shows the raw counts of one game in our dataset. But numbers do not tell the whole story. We analyze this case study using rhetorical tactics (Cialdini and Goldstein, 2004), which Oliveira et al. (2017) use to dissect spear phishing e-mails and Anand et al. (2011) apply to persuasive blogs. Mentions of tactics are in italic (e.g., *authority*. For the rest of the paper, we will refer to players via the name of their assigned country.

Through two lie-intense strategies—convincing England to betray Germany and convincing all remaining countries to agree to a draw—Italy gains control of the board. Italy's first deception is a plan with Austria to dismantle Turkey. Turkey believes Italy's initial assurance of non-aggression in 1901. Italy begins by excusing his initial silence due to a rough day at work, evoking empathy and *likability*. While they do not fall for subsequent lies, Turkey's initial gullibility cements Italy's first-strike advantage. Meanwhile, Italy proposes a long-term alliance with England against France, packaging several small truths with a big lie. The strategy succeeds, eliminating Italy's greatest threat.

Local threats eliminated, Italy turns to rivals on the other end of the map.

Italy persuades England to double-cross its long-time ally Germany in a moment of *scarcity*: if you do not act now, there will be nowhere to expand. England accepts help from ascendant Italy, expecting *reciprocity*. However, Italy aggressively and successfully moves against England. The last year features a meta-game deception. After Italy becomes too powerful to contain, the remaining four players team up. Ingeniously, Italy feigns acquiescence to a five-way draw, individually lying to each player and establishing *authority* while brokering the deal. Despite Italy's record of deception, the other players believe the proposal (annotating received messages from Italy as truthful) and expect a 1907 endgame, the year with the most lies. Italy goes on the offensive and knocks out Austria.

Each game has relationships that are forged and then riven. In another game, an honest attempt by a strong Austria to woo an ascendant Germany backfires, knocking Austria from the game. Germany builds trust with Austria through a believed fictional experience as a Boy Scout in Maine (*likability*). In a third game, two consecutive unfulfilled promises by an ambitious Russia leads to a quick demise, as their subsequent excuses and apologies are perceived as lies (failed *consistency*). In another game, England, France, and Russia simultaneously attack Germany after offering duplicitous assurances. Game outcomes vary despite the identical, balanced starting board, as different players use unique strategies to persuade, and occasionally deceive, their opponents.

7.2.2 Defining a lie

Statements can be incorrect for a host of reasons: ignorance, misunderstanding, omission, exaggeration. (Gokhman et al., 2012) highlight the difficulty of finding willful, honest, and skilled deception outside of short-term, artificial contexts (DePaulo et al., 2003). Crowdsourced and automatic datasets rely on simple negations (Pérez-Rosas et al., 2017) or completely implausible claims (e.g., "Tipper Gore was created in 1048" from (Thorne et al., 2018)). While lawyers in depositions and users of dating sites will not willingly admit to their lies, the players of online games are more willing to revel in their deception.

We must first define what we mean by deception. Lying is a mischaracterization; it's thus no surprise that a definition may be divisive or the subject of academic debate (Gettier, 1963). We provide this definition to our users: "Typically, when [someone] lies [they] say what [they] know to be false in an attempt to deceive the listener" (Siegler, 1966). An orthodox definition requires the speaker to utter an explicit falsehood (Mahon, 2016); skilled liars can deceive with a patina of veracity. A similar definition is required for prosecution of perjury, leading to a paucity of convictions (Bogner et al., 1974). Indeed, when we ask participants what a lie looks like, they mention evasiveness, shorter messages, over-qualification, and creating false hypothetical scenarios (DePaulo et al., 2003).

7.2.3 Annotating truthfulness

Previous work on the language of Diplomacy (Niculae et al., 2015) lacks access to players' internal state and was limited to *post-hoc* analysis. We improve on this by designing our own interface that gathers players' intentions and perceptions in real-time (Section 7.4.1). As with other highly subjective phenomena like sarcasm (González-Ibáñez et al., 2011; Bamman and Smith, 2015), sentiment (Pang et al., 2008) and framing (Greene and Resnik, 2009), the intention to deceive is reflective on someone's internal state. Having individuals provide their own labels for their internal state is essential as third party annotators could not accurately access it (Chang et al., 2020).

Most importantly, our gracious players have allowed this language data to be released in accordance with IRB authorized anonymization, encouraging further work on the strategic use of deception in long-lasting relations.[4]

7.3 Broader Applicability

This differs from previous work that does not follow the **expert-generated** paradigm. The most prominent past work on Diplomacy in the NLP community, (Niculae et al., 2015), did **found** their data and thus could not release it to the public. This hampers follow-up applications of the research; a believable Diplomacy-playing (and speaking) bot cannot be trained if the raw language data is redacted

[4]Data available at `http://go.umd.edu/diplomacy_data` and as part of ConvoKit `http://convokit.cornell.edu`.

and shuffled. We believe this work can set a paradigm for work outside of Diplomacy, and even NLP; the interface created for this project, as well as the pre and post-game user surveys can be modifying for any conversational task. Most importantly, building a relationship with data generators elevates the standard of the data and guarantees its liberal distribution. Further work is necessary in codifying data standards—Show Your Data, not only your Work[5].

7.4 Engaging a Community of Liars

This dataset requires both a social and technical setup: finding a community that plays Diplomacy online and having them use a framework for annotating these messages.

7.4.1 Seamless Diplomacy Data Generation

We need two technical components for our study: a game engine and a chat system. We choose Backstabbr[6] as an accessible game engine on desktop and mobile platforms: players input their moves and the site adjudicates game mechanics (Chiodini, 2020). Our communication framework is atypical. Thus, we create a server on Discord,[7] the group messaging platform most used for online gaming and by the online Diplomacy community (Coberly, 2019). The app is reliable on both desktop and mobile devices, free, and does not limit access to messages. Instead of direct

[5](Dodge et al., 2019)

[6]`https://www.backstabbr.com`

[7]`https://www.discord.com`

communication, players communicate with a bot; the bot does not forward messages to the recipient until the player annotates the messages (Figure 7.2). In addition, the bot scrapes the game state from Backstabbr to sync game and language data.

Annotation of lies is a forced binary choice in our experiment. Explicitly calling a statement a lie is difficult, and people would prefer degrees of deception (Bavelas et al., 1990; Bell and DePaulo, 1996). Thus, we follow previous work that views linguistic deception as binary (Buller et al., 1996; Braun and Van Swol, 2016). Some studies make a more fine-grained distinction; for example, Swol et al. (2012) separate strategic omissions from blatant lies (we consider both deception). However, because we are asking the speakers themselves (and not trained annotators) to make the decision, we follow the advice from crowdsourcing to simplify the task as much as possible (Snow et al., 2008; Sabou et al., 2014). Long messages can contain both truths and lies, and we ask players to categorize these as lies since the truth can be a shroud for their aims.

7.4.2 Building a player base

The Diplomacy players maintain an active, vibrant community through real-life meetups and online play (Hill, 2014; Chiodini, 2020). We recruit top players alongside inexperienced but committed players in the interest of having a diverse pool. Our experiments include top-ranked players and community leaders from online platforms, grizzled in-person tournament players with over 100 past games, and board game aficionados. These players serve as our foundation and during

initial design helped us to create a minimally annoying interface and a definition of a lie that would be consistent with Diplomacy play. Good players—as determined by active participation, annotation and game outcome—are asked to play in future games.

In traditional crowdsourcing tasks compensation is tied to piecework that takes seconds to complete (Buhrmester et al., 2011). Diplomacy games are different in that they can last a month. . . and people already play the game for free. Thus, we do not want compensation to interfere with what these players already do well: lying. Even the obituary of the game's inventor explains

> Diplomacy rewards all manner of mendacity: spying, lying, bribery, rumor mongering, psychological manipulation, outright intimidation, betrayal, vengeance and backstabbing (the use of actual cutlery is discouraged)" (Fox, 2013).

Thus, our goal is to have compensation mechanisms that get people to play this game as they normally would, finish their games, and put up with our (slightly) cumbersome interface. Part of the compensation is non-monetary: a game experience with players that are more engaged than the average online player.

To encourage complete games, most of the payment is conditioned on finishing a game, with rewards for doing well in the game. Players get at least $40 upon finishing a game. Additionally, we provide bonuses for specific outcomes: $24 for winning the game (an evenly divisible amount that can be split among remaining players) and $10 for having the most successful lies, i.e., statements they marked

as a lie that others believed.[8] Diplomacy usually ends with a handful of players dividing the board among themselves and agreeing to a tie. In the game described in Section 7.2.1, the remaining four players shared the winner's pool with Italy after 10 in-game years, and Italy won the prize for most successful lies.

7.4.3 Data overview

Table 7.2 quantitatively summarizes our data. Messages vary in length and can be paragraphs long (Figure 7.3). Close to five percent of all messages in the dataset are marked as lies and almost the same percentage (but not necessarily the same messages) are perceived as lies, consistent with the "veracity effect" (Levine et al., 1999). In the game discussed above, eight percent of messages are marked as lies by the sender and three percent of messages are perceived as lies by the recipient; however, the messages perceived as lies are rarely lies (Figure 7.4).

7.4.4 Demographics and self-assessment

We collect anonymous demographic information from our study participants: the average player identifies as male, between 20 and 35 years old, speaks English as their primary language, and has played over fifty Diplomacy games.[9] Players

[8]The lie incentive is relatively small (compared to incentives for participation and winning) to discourage an opportunistic player from marking everything as a lie. Games were monitored in real-time and no player was found abusing the system (marking more than ∼20% lies).

[9]Our data skews 80% male and 95% of the players speak English as a primary language. Ages range from eighteen and sixty-four. Game experience is distributed across beginner, intermediate, and expert levels.

self-assess their lying ability before the study. The average player views themselves as better than average at lying and average or better than average at perceiving lies.

In a post-game survey, players provide information on whom *they* betrayed and who betrayed *them* in a given game. This is a finer-grained determination than the *post hoc* analysis used in past work on Diplomacy (Niculae et al., 2015). We ask players to optionally provide linguistic cues to their lying and to summarize the game from their perspective.

7.4.5 An ontology of deception

Four possible combinations of deception and perception can arise from our data. The sender can be lying or telling the truth. Additionally, the receiver can perceive the message as deceptive or truthful. We name the possible outcomes for lies as Deceived or Caught, and the outcomes for truthful messages as Straightforward or Cassandra,[10] based on the receiver's annotation (examples in Table 7.3, distribution in Figure 7.4).

7.5 Detecting Lies

We build computational models both to detect lies to better understand our dataset. The data from the user study provide a training corpus that maps language to annotations of truthfulness and deception. Our models progressively integrate information—conversational context and in-game power dynamics—to approach hu-

[10] In myth, Cassandra was cursed to utter true prophecies but never be believed. For a discussion of Cassandra's curse *vis a vis* personal and political oaths, see Torrance (2015).

man parity in deception detection.

7.5.1 Metric and data splits

We investigate two phenomena: detecting what is *intended* as a lie and what is *perceived* as a lie. However, this is complicated because most statements are not lies: less than five percent of the messages are labeled as lies in both the ACTUAL LIE and the SUSPECTED LIE tasks (Table 7.2). Our results use a weighted F_1 feature across truth and lie prediction, as accuracy is an inflated metric given the class imbalance (Japkowicz and Stephen, 2002). We thus adopt an in-training approach (Zhou and Liu, 2005) where incorrect predictions of lies are penalized more than truthful statements. The relative penalty between the two classes is a hyper-parameter tuned on F_1.

Before we move to computational models for lie detection, we first establish the *human* baseline. We know when senders were lying and when receivers spotted a lie. Humans spot 88.3% of lies. However, given the class imbalance, this sounds better than it is. Following the suggestion of (Levine et al., 1999), we focus on the detection of lies, where humans have a 22.5 Lie F_1.

To prevent overfitting to specific games, nine games are used as training data, one is used for validation for tuning parameters, and two games are test data. Some players repeat between games.

7.5.2 Logistic regression

Logistic regression models (Background Section **??**) have interpretable coefficients which show linguistic phenomena that correlate with lies. A *word* that occurs infrequently overall but often in lies, such as 'honest' and 'candidly', helps identify which messages are lies.

(Niculae et al., 2015) propose linguistic **Harbingers** that can predict deception. These are word lists that cover topics often used in interpersonal communication—*claims*, *subjectivity*, *premises*, *contingency*, *comparisons*, *expansion*, *temporal language associated with the future*, and *all other temporal language*. The Harbingers word lists do not provide full coverage, as they focus on specific rhetorical areas. A logistic regression model with all word types as features further improves F_1.

Power dynamics influence the language and flow of conversation (Danescu-Niculescu-Mizil et al., 2012, 2013; Prabhakaran et al., 2013). These dynamics may influence the likeliness of lying; a stronger player may feel empowered to lie to their neighbor. Recall that victory points (Section 7.2) encode how well a player is doing (more is better). We represent the power differential as the difference between the two players. Peers will have a zero differential, while more powerful players will have a positive differential with their interlocutor. The differential changes throughout the game, so this feature encodes the difference in the season the message was sent. For example, a message sent by an Italy with seven points to a Germany with two points in a given season would have a value of five.

7.5.3 Neural

While less interpretable, neural models are often more accurate than logistic regression ones (Ribeiro et al., 2016; Belinkov and Glass, 2019). We build a standard long short-term memory network (Hochreiter and Schmidhuber, 1997, LSTM) (Background Section **??**) to investigate if word sequences—ignored by logistic regression—can reveal lies.

Integrating message context and power dynamics improves on the neural baseline. A Hierarchical LSTM can help focus attention on specific phrases in long conversational contexts. In the same way it would be difficult for a human to determine *prima facie* if a statement is a lie without previous context, we posit that methods that operate at the level of a single message are limited in the types of cues they can extract. The hierarchical LSTM is given the context of previous messages when determining if a given message is a lie, which is akin to the labeling task humans do when annotating the data. The model does this by encoding a single message from the tokens, and then running a forward LSTM over all the messages. For each message, it looks at both the content and previous context to decide if the current message is a lie. Fine-tuning BERT (Devlin et al., 2019) embeddings, introduced in Background Section **??**, to this model did not lead to notable improvement in F_1, likely due to the relative small size of our training data. Last, we incorporate information about power imbalance into this model. This model approaches human performance in terms of F_1 score by combining content with conversational context and power imbalance.

7.6 Qualitative Analysis

This section examines specific messages where both players and machines are correctly identifying lies and when they make mistakes on our test set. Most messages are correctly predicted by both the model and players (2055 of 2475 messages); but this is because of the veracity effect. The picture is less rosy if we only look at messages the sender marks as ACTUAL LIE: both players and models are generally wrong (Table 7.5).

Both models and players can detect lies when liars get into specifics. In Diplomacy, users must agree to help one another through orders that stipulate "I will help another player move from X to Y". The in-game term for this is "support"; half the messages where players and computers correctly identify lies contain this word, but it rarely occurs in the other quadrants.

Models seem to be better at not falling for vague excuses or fantastical promises in the future. Players miss lies that promise long-term alliances, involve extensive apologies, or attribute motivation as coming from other countries' disinformation (*Model Correct*). Unlike our models, players have access to conversations with other players and accordingly players can detect lies that can easily be verified through conversations with other players (*Player Correct*).

However, ultimately most lies are believable and fool both models and players (*Both Wrong*). For example, all messages that contain the word "true" are predicted as truthful by both models and players. Many of these messages are rel-

atively tame;[11] confirming the Pinocchio effect found by Swol et al. (2012). If liars can be detected when they wax prolix, perhaps the best way to avoid detection is to be terse and to the point.

Sometimes additional contextual information helps models improve over player predictions. For example, when France tells Austria "I am worried about a steam-roller Russia Turkey alliance", the message is incorrectly perceived as truthful by both the player and the single-message model. However, once the model has context—a preceding question asking if Austria and Turkey were cooperating—it can detect the lie.

Finally, we investigate categories from the Harbingers (Niculae et al., 2015) word lists. Lies are more likely to contain *subjectivity* and *premises* while true messages include *expansion* phrases ("later", "additionally"). We also use specific words in the bag of words logistic regression model. The coefficient weights of words that express sincerity (e.g., "sincerely", "frankly") and apology (e.g., "accusation", "fallout", "alternatives") skew toward ACTUAL LIE prediction in the logistic regression model. More laid back appellations (e.g., "dude", "man") skew towards truthfulness, as do words associated with reconnaissance (e.g., "fyi","useful", "information") and time (e.g., "weekend", "morning"). Contested areas on the Diplomacy map, such as Budapest and Sevastopol, are more likely to be associated with lies, while more secure ones like Berlin, are more likely to be associated with truthful messages.

[11] Examples include "It's true—[Budapest] back to [Rumania] and [Serbia] on to [Albania] could position for more forward convoys without needing the rear fleet..." and "idk if it's true just letting u know since were allies".

7.7 Related Work

Early computational deception work focuses on single utterances (Newman et al., 2003), especially for product reviews (Ott et al., 2012). But deception is intrinsically a discursive phenomenon and thus the context in which it appears is essential. Our platform provides an opportunity to observe deception in the context in which it arises: goal-oriented conversations around in-game objectives. Gathering data through an interactive game has a cheaper per-lie cost than hiring workers to write deceptive statements (Jurgens and Navigli, 2014).

Other conversational datasets are mostly based on games that involve deception including Werewolf (Girlea et al., 2016), Box of Lies (Soldner et al., 2019), and tailor-made games (Ho et al., 2017). However, these games assign individuals roles that they maintain throughout the game (i.e., in a role that is supposed to deceive or in a role that is deceived). Thus, deception labels are coarse: an *individual* always lies or always tells the truth. In contrast, our platform better captures a more multi-faceted reality about human nature: everyone can lie or be truthful with everyone else, and they use both strategically. Hence, players must think about *every* player lying at any moment: "given the evidence, do I think this person is lying to me *now*?"

Deception data with conversational labels is also available through interviews (Pérez-Rosas et al., 2016), some of which allow for finer-grained deception spans (Levitan et al., 2018). Compared with game-sourced data, however, interviews provide shorter conversational context (often only a single exchange with a few follow-ups)

and lack a strategic incentive—individuals lie because they are instructed to do so, not to strategically accomplish a larger goal. In Diplomacy, users have an intrinsic motivation to lie; they have entertainment-based and financial motivations to win the game. This leads to higher-quality, creative lies.

Real-world examples of lying include perjury (Louwerse et al., 2010), calumny (Fornaciari and Poesio, 2013), emails from malicious hackers (Dhamija et al., 2006), and surreptitious user recordings. But real-world data comes with real-world complications and privacy concerns. The artifice of Diplomacy allows us to gather pertinent language data with minimal risk and to access both sides of deception: intention and perception. Other avenues for less secure research include analyzing dating profiles for accuracy in self-presentation (Toma and Hancock, 2012) and classifying deceptive online spam (Ott et al., 2011).

7.8 Detecting Deception

In Dante's *Inferno*, the ninth circle of Hell—a fate worse even than that reserved for murderers—is for betrayers. Dante asks Count Ugolino to name his betrayer, which leads him to say:

> but if my words can be the seed to bear
> the fruit of infamy for this betrayer
> who feeds my hunger, then I shall speak—in tears (Alighieri and Musa, 1995, Canto XXXIII)

Similarly, we ask victims to expose their betrayers in the game of Diplomacy. The

seeds of players' negotiations and deceit could, we hope, yield fruit to help others: understanding multi-party negotiation and protecting Internet users.

While we ignore nuances of the game board to keep our work general, Diplomacy is also a rich, multi-agent strategic environment; (Paquette et al., 2019) ignore Diplomacy's rich language to build bots that only move pieces around the board. An exciting synthesis would incorporate deception and language generation into an agent's policy; our data would help train such agents. Beyond playing against humans, playing with a human in the loop (HITL) resembles designs for cybersecurity threats (Cranor, 2008), annotation (Branson et al., 2010), and language alteration (Wallace et al., 2019). Likewise, our lie-detection models can help a user *in the moment* better decide whether they are being deceived (Lai et al., 2020). Computers can meld their attention to detail and nigh infinite memory to humans' grasp of social interactions and nuance to forge a more discerning player.

Beyond a silly board game, humans often need help verifying claims are true when evaluating health information (Xie and Bugg, 2009), knowing when to take an e-mail at face value (Jagatic et al., 2007), or evaluating breaking news (Hassan et al., 2017). Building systems to help information consumers become more discerning and suspicious in low-stakes settings like online Diplomacy are the seeds that will bear the fruits of interfaces and machine learning tools necessary for a safer and more robust Internet ecosystem.

7.9 The Expert Edge

This dataset is created by expert users, in this case Diplomacy players, In contrast to CANARD (Chapter **??**). While there are quality differences even within a verified pool of community-of-interest, only one out of 80 users did not actively participate in the experiment. In contrast, dozens of workers had be blocked due to noncompliance during the collection process for CANARD. The data are thoughtful, clever, and sometimes even funny, which are adjectives that seldom apply to large-scale NLP datasets.

Experts from other areas would be necessary to extend this work to conversational areas of NLP broader than Diplomacy. **Generation** and **annotation** of conversational Diplomacy data would not be possible without expertise, given the lexicon of the game. A challenge of extending and scaling our approach to data collection is that this community of interest is limited in size, so data collection took a year. Chapter **??** explores a hybrid approach that pairs **experts** with the **crowd** to **generate** a larger and more general conversational dataset.

Message	**Sender's intention**	**Receiver's percep.**
If I were lying to you, I'd smile and say "that sounds great." I'm honest with you because I sincerely thought of us as partners.	Lie	Truth
You agreed to warn me of unexpected moves, then didn't . . . You've revealed things to England without my permission, and then made up a story about it after the fact!	Truth	Truth
. . . I have a reputation in this hobby for being sincere. Not being duplicitous. It has always served me well. . . . If you don't want to work with me, then I can understand that . . .	Lie	Truth
(Germany attacks Italy)		
Well this game just got less fun	Truth	Truth
For you, maybe	Truth	Truth

Table 7.1: An annotated conversation between Italy (white) and Germany (gray) at a moment when their relationship breaks down. Each message is annotated by the sender (and receiver) with its intended or perceived truthfulness; Italy is lying about . . . lying.

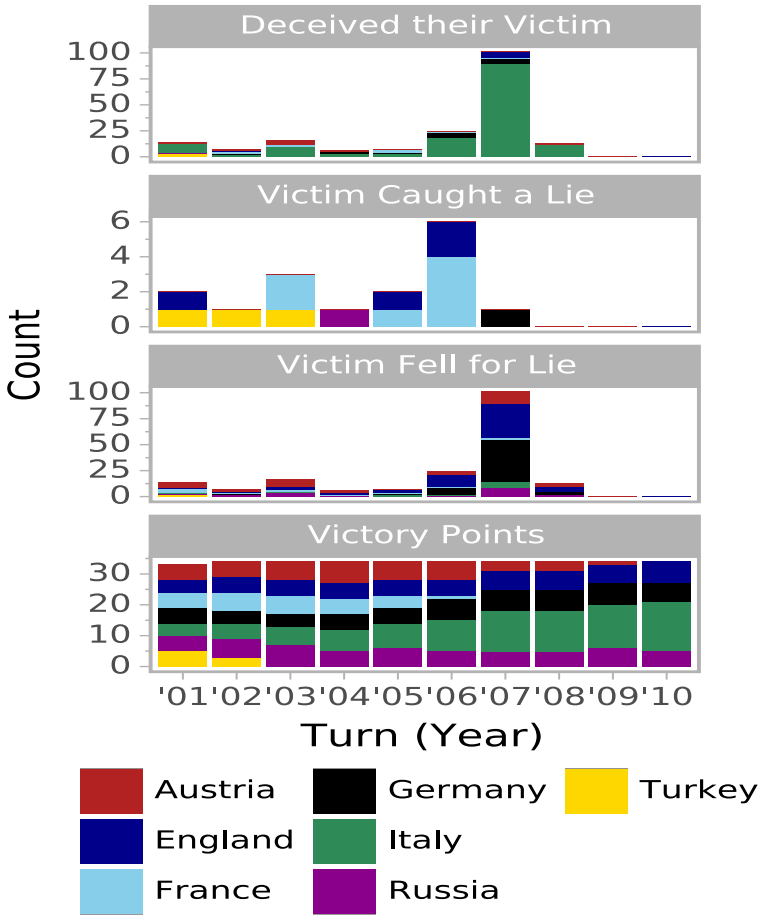

Figure 7.1: Counts from one game featuring an Italy (green) adept at lying but who does not fall for others' lies. The player's successful lies allow them to gain an advantage in points over the duration of the game. In 1906, Italy lies to England before breaking their relationship. In 1907, Italy lies to everybody else about wanting to agree to a draw, leading to the large spike in successful lies.

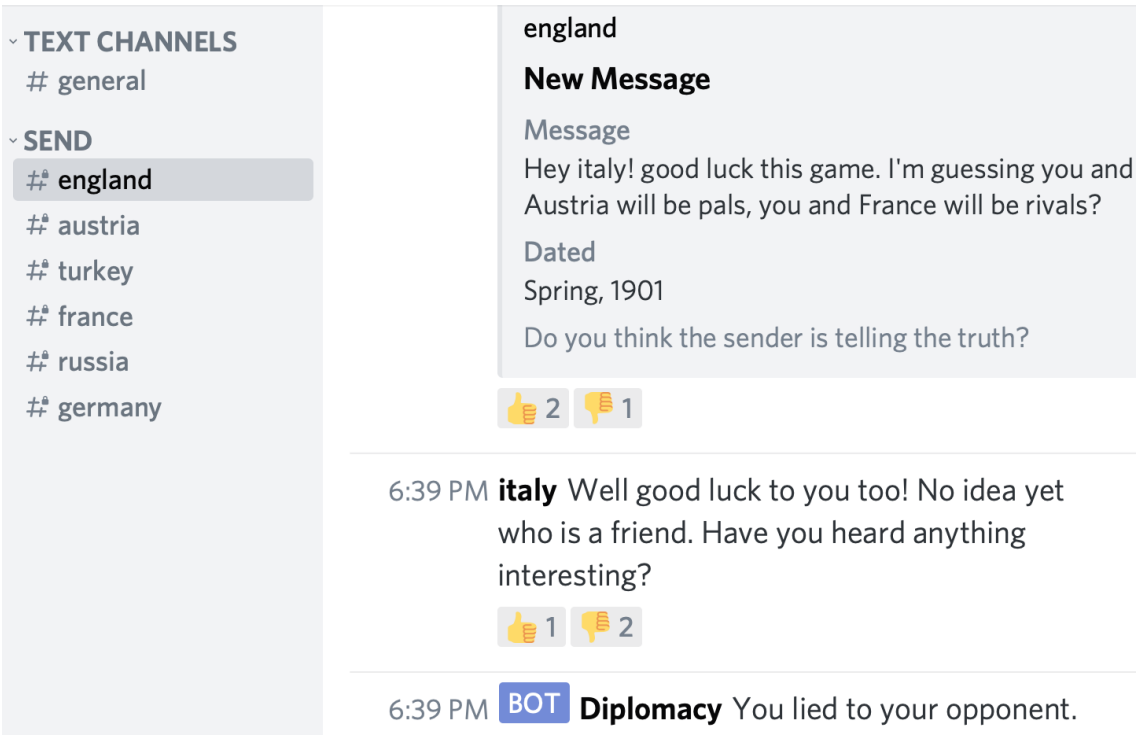

Figure 7.2: Every time they send a message, players say whether the message is truthful or intended to deceive. The receiver then labels whether incoming messages are a lie or not. Here Italy indicates they believe a message from England is truthful but that their reply is not.

Category	Value
Message Count	13,132
ACTUAL LIE Count	591
SUSPECTED LIE Count	566
Average # of Words	20.79

Table 7.2: Summary statistics for our train data (nine of twelve games). Messages are long and only five percent are lies, creating a class imbalance.

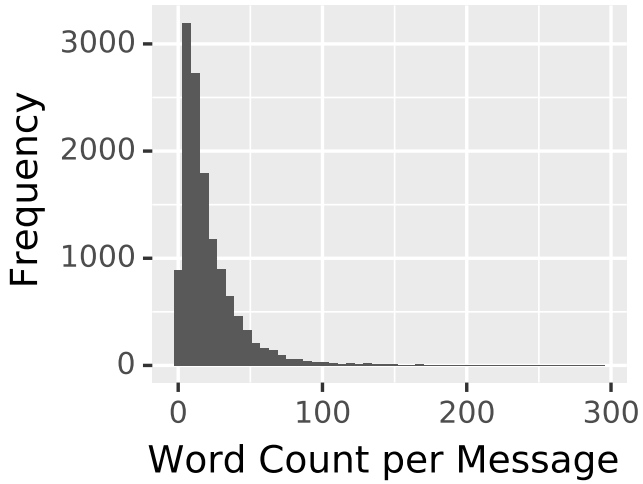

Figure 7.3: Individual messages can be quite long, wrapping deception in pleasantries and obfuscation.

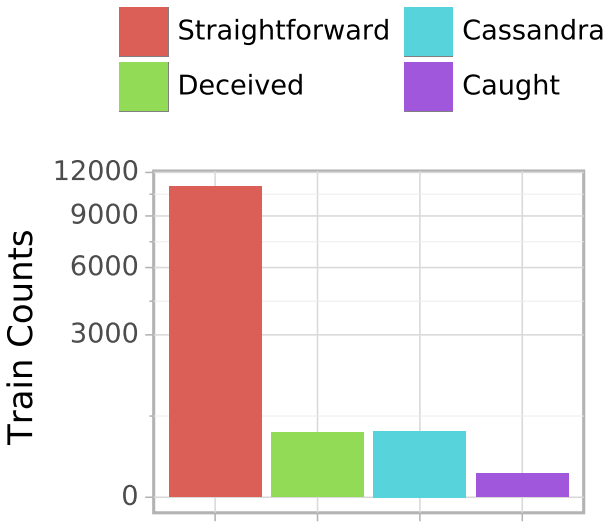

Figure 7.4: Most messages are truthful messages identified as the truth. Lies are often not caught. Table 7.3 provides an example from each quadrant.

		Receiver's perception	
		Truth	**Lie**
Sender's intention	**Truth**	**Straightforward** Salut! Just checking in, letting you know the embassy is open, and if you decide to move in a direction I might be able to get involved in, we can probably come to a reasonable arrangement on cooperation. Bonne journee!	**Cassandra** I don't care if we target T first or A first. I'll let you decide. But I want to work as your partner. ...I literally will not message anyone else until you and I have a plan. I want it to be clear to you that you're the ally I want.
	Lie	**Deceived** You, sir, are a terrific ally. This was more than you needed to do, but makes me feel like this is really a long term thing! Thank you.	**Caught** So, is it worth us having a discussion this turn? I sincerely wanted to work something out with you last turn, but I took silence to be an ominous sign.

Table 7.3: Examples of messages that were intended to be truthful or deceptive by the sender or receiver. Most messages occur in the top left quadrant (Straightforward). Figure 7.4 shows the full distribution. Both the intended and perceived properties of lies are of interest in our study.

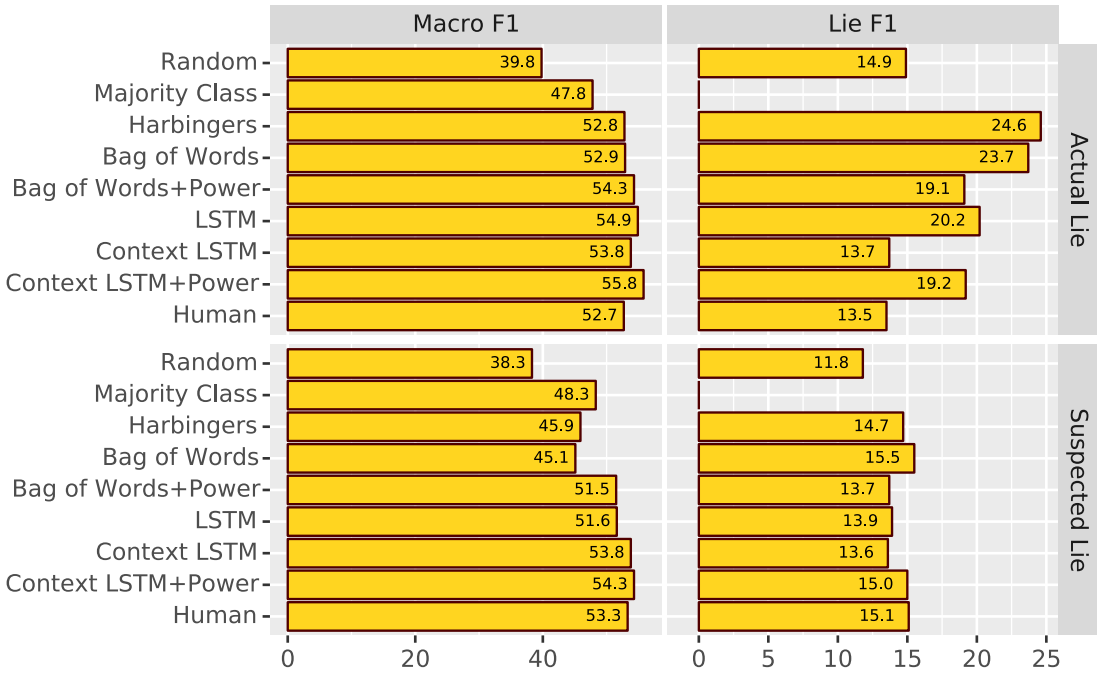

Figure 7.5: Test set results for both our ACTUAL LIE and SUSPECTED LIE tasks. We provide baseline (Random, Majority Class), logistic (language features, bag of words), and neural (combinations of a LSTM with BERT) models. The neural model that integrates past messages and power dynamics approaches human F_1 for ACTUAL LIE (top). For ACTUAL LIE, the human baseline is how often the receiver correctly detects senders' lies. The SUSPECTED LIE lacks such a baseline.

		Model Prediction: Correct	Model Prediction: Wrong
Player Prediction	**Correct**	**Both Correct** Not sure what your plan is, but I might be able to support you to Munich.	**Player Correct** Don't believe Turkey, I said nothing of the sort. I imagine he's just trying to cause an upset between us.
	Wrong	**Model Correct** Long time no see. Sorry for the stab earlier. I think we should try to work together to stop france from winning; if we work together we can stop france from getting 3 more centers, and then we will all win in a 3, 4, or 5 way draw when the game is hard-capped at 1910.	**Both Wrong** I'm considering playing fairly aggressive against England and cutting them off at the pass in 1901, your support for that would be very helpful.

Table 7.4: An example of an ACTUAL LIE detected (or not) by both players and our best computational model (Context LSTM + Power) from each quadrant. Both the model and the human recipient are mostly correct overall (Both Correct), but they are both mostly wrong when it comes to specifically predicting lies (Both Wrong).

	Model Correct	**Model Wrong**
Player Correct	10	32
Player Wrong	28	137

Table 7.5: Conditioning on only lies, most messages are now identified incorrectly by both our best model (Context LSTM + Power) and players.

Bibliography

Dante Alighieri and Mark Musa. 1995. *Dante's Inferno: The Indiana Critical Edition*. Indiana masterpiece editions. Indiana University Press.

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Douglas W. Oard, and Philip Resnik. 2011. Believe me: We can do this! In *The AAAI 2011 workshop on Computational Models of Natural Argument*.

David Bamman and Noah A. Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Proceedings of ICWSM*.

Janet Beavin Bavelas, Alex Black, Nicole Chovil, and Jennifer Mullett. 1990. Truths, lies, and equivocations: The effects of conflicting goals on discourse. *Journal of Language and Social Psychology*, 9(1-2):135–161.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*.

Kathy L Bell and Bella M DePaulo. 1996. Liking and lying. *Basic and Applied Social Psychology*, 18(3):243–266.

William E. Bogner, Margaret Edwards, Leon Zelechowski, Kevin J. Egan, William J. Rogers, Eloy Burciaga, and John Scott Arthur. 1974. Perjury: The forgotten offense. *The Journal of Criminal Law and Criminology*, 65(3):361–372.

Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual recognition with humans in the loop. In *European Conference on Computer Vision*.

Michael T. Braun and Lyn M. Van Swol. 2016. Justifications offered, questions asked, and linguistic patterns in deceptive and truthful monetary interactions. *Group Decision and Negotiation*, 25(3):641–661.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on psychological science: a journal of the Association for Psychological Science*, 6 1:3–5.

David B. Buller, Judee K. Burgoon, Aileen Buslig, and James Roiger. 1996. Testing interpersonal deception theory: The language of interpersonal deception. *Communication Theory*, 6(3):268–289.

Jonathan P. Chang, Justin Cheng, and Cristian Danescu-Niculescu-Mizil. 2020. Don't let me be misunderstood: Comparing intentions and perceptions in online discussions. In *Proceedings of the World Wide Web Conference*.

Johnny Chiodini. 2020. Playing Diplomacy online transformed the infamously brutal board game from unbearable to brilliant. *Dicebreaker*.

Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621.

Cohen Coberly. 2019. Discord has surpassed 250 million registered users. *Techspot*.

B. Cornwell and D. C. Lundgren. 2001. Love on the internet: involvement and misrepresentation in romantic relationships in cyberspace vs. realspace. *Computational Human Behavior*, 17:197–211.

Lorrie F Cranor. 2008. A framework for reasoning about the human in the loop. In *UPSEC*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the World Wide Web Conference*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the Association for Computational Linguistics*.

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Rachna Dhamija, J. Doug Tygar, and Marti A. Hearst. 2006. Why phishing works. In *International Conference on Human Factors in Computing Systems*.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of Empirical Methods in Natural Language Processing*.

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

Margalit Fox. 2013. Allan Calhamer dies at 81; invented Diplomacy game. *New York Times*.

Edmund Gettier. 1963. Is justified true belief knowledge? *Analysis*, 23(6):121–123.

Codruta Girlea, Roxana Girju, and Eyal Amir. 2016. Psycholinguistic features for deceptive role detection in Werewolf. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Stephanie Gokhman, Jeff Hancock, Poornima Prabhu, Myle Ott, and Claire Cardie. 2012. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the Association for Computational Linguistics*.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Knowledge Discovery and Data Mining*.

David Hill. 2014. Got your back. *This American Life Podcast*.

Shuyuan Mary Ho, Jeffrey T Hancock, and Cheryl Booth. 2017. Ethical dilemma: Deception dynamics in computer-mediated group communication. *Journal of the Association for Information Science and Technology*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. *Communications of the ACM*, 50(10):94–100.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. In *Transactions of the Association for Computational Linguistics*.

Mary E Kaplar and Anne K Gordon. 2004. The enigma of altruistic lying: Perspective differences in what motivates and justifies lie telling within romantic relationships. *Personal Relationships*, 11(4):489–507.

Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *Proceedings of the World Wide Web Conference*, Republic and Canton of Geneva, Switzerland.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. In *Social Media Analytics: Advances and Applications*. CRC.

Vivian Lai, Han Liu, and Chenhao Tan. 2020. "why is 'chicago' deceptive?" Towards building model-driven tutorials for humans. In *International Conference on Human Factors in Computing Systems*.

Timothy R. Levine, Hee Sun Park, and Steven A. McCornack. 1999. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs*, 66(2):125–144.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. Linguistic cues to deception and perceived deception in interview dialogues. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Max Louwerse, David Lin, Amanda Drescher, and Gun Semin. 2010. Linguistic cues predict fraudulent events in a corporate social network. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

James Edwin Mahon. 2016. The definition of lying and deception. In *The Stanford Encyclopedia of Philosophy*, winter 2016 edition. Metaphysics Research Lab, Stanford University.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the Association for Computational Linguistics*.

Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. 2017. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *International Conference on Human Factors in Computing Systems*.

Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the World Wide Web Conference*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the Association for Computational Linguistics*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Philip Paquette, Yuchen Lu, Seton Steven Bocco, Max Smith, Satya Ortiz-Gagné, Jonathan K. Kummerfeld, Joelle Pineau, Satinder Singh, and Aaron C Courville. 2019. No-press diplomacy: Modeling multi-agent gameplay. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, C. J. Linton, and Mihai Burzo. 2016. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of Empirical Methods in Natural Language Processing*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *Proceedings of International Conference on Computational Linguistics*.

Vinodkumar Prabhakaran, Ajita John, and Dorée D Seligmann. 2013. Power dynamics in spoken interactions: a case study on 2012 Republican primary debates. In *Proceedings of the World Wide Web Conference*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Language Resources and Evaluation Conference*.

Ben Shneiderman. 2000. Designing trust into online experiences. *Communications of the ACM*, 43(12):57–59.

Frederick A Siegler. 1966. Lying. *American Philosophical Quarterly*, 3(2):128–136.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Lyn M. Van Swol, Deepak Malhotra, and Michael T. Braun. 2012. Deception and its detection: Effects of monetary incentives and personal relationship history. *Communication Research*, 39(2):217–238.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal, editors. 2018. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Catalina L Toma and Jeffrey T Hancock. 2012. What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62(1):78–97.

Isabelle Torrance. 2015. Distorted oaths in Aeschylus. *Illinois Classical Studies*, 40(2):281–295.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Bo Xie and Julie M. Bugg. 2009. Public library computer training for older adults to access high-quality internet health information. *Library and Information Science Research*, 31(3).

Zhi-Hua Zhou and Xu-Ying Liu. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77.