2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)

Decoding the Perception of Sincerity in Written Dialogues

Codruta Girlea
University of Illinois at Urbana-Champaign
Email: codruta.liliana@gmail.com

Roxana Girju
University of Illinois at Urbana-Champaign
Email: girju@illinois.edu

Abstract—**Understanding how people perceive sincerity is important in identifying malevolent deceptive behavior in social media. This is a difficult problem because the only signal in most such cases is written text, and the larger proportion of human communication is non-verbal. Furthermore, perception is very subjective, and people adapt to each other's perceptions.**

We define the problem of decoding sincerity perception for two domains, dating and online games. We introduce and evaluate new psycholinguistic features of perceived sincerity. Our results shed light into the connection between language, deception, and perception, and underline the challenges and difficulty of assessing perceptions from written text.

1. Introduction

Technological progress has made it commonplace for people to connect across large distances. In environments such as social media and Internet forums, most often the only signal is written language. Since the verbal component of communication is very small [1], intentions become easier to hide and harder to read. This increases the risk of harmful interactions, such as bullying, trolling, or predatory behavior.

The harmful aspect of such malevolent communication is that it is deceptive. The manifest intent of the speaker is different from the speaker's true intent. For example, the manifest intent of trolling [2], [3] is to discuss arguments that the speaker genuinely believes, whereas the true intent is to annoy interlocutors and boost the troll's self-worth.

Deception detection in written dialogues is a difficult task. Deception has long been one of the least studied areas in natural language processing. Furthermore, only 7% of human communication is verbal [1], while over 90% is comprised of tone of voice (38%) and body language (55%) (the 7% Rule). Because of this, written dialogues are very difficult to process, since most of the non-verbal communication is either missing or needs to be inferred.

We propose that the task of understanding deception involves understanding both when a speaker intends to deceive, and when a listener perceives a speaker as sincere. Deception happens when a *deceptive* speaker is *perceived as sincere*. In this paper, we focus on the second task.

While deception in dialogues has been somewhat explored [4], [5], [6], [7], perceived sincerity has only very recently gained attention [8], [9]. However, the focus in the work to date has been on audio or prosodic features.

Here, we aim to address the following questions: 1) How hard is the problem of decoding perceived sincerity with no or poor paralinguistic context? and 2) Which psycholinguistic features are helpful in assessing perceived sincerity in two challenging written dialog settings: games and dating?

We contribute an analysis and experimental evaluation of our novel psycholinguistic features of perceived sincerity.

We will focus on two very different domains: dating [10] and online text dialog games such as Werewolf [6]. Deception is an important phenomenon in dating [11], [12] and understanding the perception of sincerity is helpful in assessing its impact. Games like Werewolf are designed in such a way that people are motivated to deceive and detect deception in order to win, which makes them useful in understanding human behavior.

For dating, the gold standard is self-reported perceived sincerity of interlocutor after 4 minutes of dialogue. For Werewolf, we have no self-reported gold standard. We assign scores based on the extent to which other players agree, through voting, on a player's deceptive role.

2. Related Work

Most research on deception in written language focuses on non-interactive media such as opinions [13], product reviews [14], or dating profiles [15]. The focus of our work is on the interactive setting of dialogues.

Research in computational linguistics on deception in dialogues considered the domain of interviews [4], [5] or games like Werewolf [6] and Diplomacy [7] (see [16]) for a summary of work in deception detection). Such games require players to discuss and use deception in order to win. Some approaches focus on written language only, whereas others [6] use audio data as well. Our focus is on the related but different problem of perception-of-sincerity.

A recent paralinguistic challenge [8] has drawn attention to the problem of deception and sincerity in dialogues. The challenge and the winning models for sincerity [17] focus on acoustic signal. In our setting, we are interested in modeling perceived sincerity using written language.

There has been research on deception and sincerity in conversations, in psycholinguistics [9], [18]. However, for deception and perception-of-deception in dialogues [9], the data contains acoustic information as well, and most features refer to prosody, disfluency, and vocal characteristics.

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

Work on mapping psychological dimensions to word categories [19] has identified several classes of words for deception. This work does not focus on dialogue and does not tackle deception in particular.

For the problem of identifying stances in dialogues [20], [21], several such word categories as well as discourse features were shown to perform better than the human baseline (human perception) in the dating domain. Deception or its perception were however not considered among the stances.

Our work draws attention to the task of decoding perceived sincerity in an interactive setting where written language is the only input. We introduce new features that are helpful for this task. We also show that the problem, while important, is much more difficult than in the multimodal case of face-to-face dialogue.

3. Data

We first focus on the domain of dating, where establishing trust is important and people might be tempted to misrepresent themselves. We are using the SpeedDate corpus [10]. Next, we focus on games in which deception is necessary, in particular Werewolf [6].

1) Speed Dating

The SpeedDate dataset has been used before to detect flirting [20] and other interpersonal stances [21]. It consists of 1100 recordings of anonymized 4-minute dates.

We are using the transcriptions of those recordings. Many paralinguistic features are not present in the transcriptions. However, each utterance is marked with the beginning and end timestamps. Laughter is also marked.

There are 163 dialogue participants, 54 female and 109 male. Each conversation happens between a female and a male and has up to 331 utterances (123.14 on average).

The dialogue participants have also been asked to provide the following information:
(i) **(met)** – how well they know the partner (1-6)
(ii) **(willng)** – willingness to give contact information.
(iii) **(timemtch)** – the time it took to decide whether to select that partner or not (1-5 minutes, or 6 for later).
(iv) **self stance**, of the person reporting – how often (1-10) was the person friendly (s-fndly), flirtatious (s-flirt), awkward (s-awk), or assertive (s-assert).
(v) **other's stance**, or the perceived stance of the interlocutor – how often (1-10) was the other person friendly (o-fndly), flirtatious (o-flirt), awkward (o-awk), or assertive (o-assert).
(vi) **other's qualities**, or the perceived qualities of the interlocutor, on a 1-10 scale – how attractive (o-attrct), sincere (o-sincre), intelligent (o-intell), funny (o-funny), ambitious (o-ambits), or courteous (o-crteos) was the other person.

We expect that trustworthiness plays a large part in whether a person is willing to give contact information **(willng)**. In the next subsection, we test this by evaluating the extent to which the other labels can predict **(willng)**.

Since the task is to classify perceived sincerity, we use self-reported perceived sincerity as a label (in others' qualities – **o-sincre**). The problem is to identify the extent to which the speaker is perceived by the listener as sincere. We consider both the multi-class and the binary classification problem. The label distribution is shown in Figure 1. The most balanced binarization is for threshold T=3: [1 – 3] positive and [4 – 10] negative (54.06% negative).

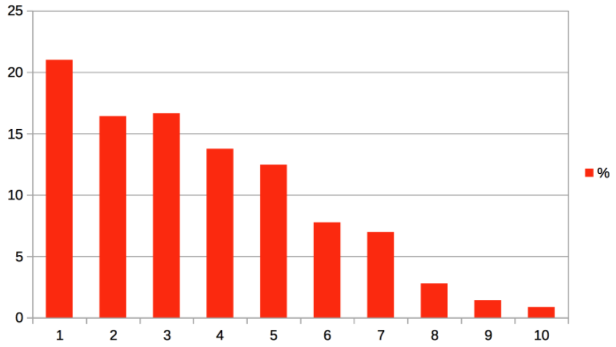

Figure 1. Label distribution for perceived sincerity in SpeedDate

We also performed significance testing to evaluate the extent to which other labels correlate with the perceived sincerity label. We did not find significant correlations.

Willingness to give contact information in speed dating

Our intuition is that giving out the e-mail **(willng)** is a statement of trust. We therefore expect perceived sincerity **(o-sincre)** to be a strong predictor of **willng**.

We evaluated the relationship between various labels in the dataset and **willng**. We use logistic regression (LR) and support vector machines (SVM) as classifiers, with 10-fold cross validation (Table 1). We use subsets of the other labels as features: self stances (e.g. s-assert), other's perceived stances (e.g. o-assert), and self stances as perceived by the other (e.g. os-assert). We also show significant correlations of other labels to **(willng)** in Table 2. Willingness to give out the e-mail is a binary variable (45.31% positive responses).

Perceived sincerity: As Tables 1 and 2 show, perceived sincerity **(o-sincre)** has little influence on deciding whether someone is willing to give out their e-mail address.

Perceived stances and flirting: As expected, perceived stances **(o-stances)** have more influence on the decision than self-reported stances **(os-stances)**. This holds for each stance individually **(assert, awk, fndly, flirt)**, as well as all together. However, the strongest influence is **flirting**.

On flirting, we note that whether or not one is flirting **(s-flirt)** has a greater influence on one's willingness, than whether they are perceived as flirting or whether the other is flirting. This makes sense given the common cause: if someone wants to connect, they may be more likely to flirt.

Time it takes to decide: As expected, the time it takes to make the decision **(timemtch)** is negatively correlated to the willingness to give out the e-mail (Table 2).

Qualities of the interlocutor: The qualities that do influence the decision on whether to give out e-mail or not are attractiveness **(o-attrct)** and humor **(o-funny)**.

Conclusion: The strongest predictors of the willingness variable seem to have little to do with how honest the inter-

TABLE 1. 10-fold cross validation accuracy of classifying **willng** based on various subsets of outcomes. Bolded are the best feature sets. Shorthand notation used: stances = {fndly, flirt, assert, awk}; o-all = {o-fndly, o-flirt, o-assert, o-awk, o-attrct, o-intell, o-funny, o-ambits, o-crteous, o-sincre}; tm = time to decide; X - {o-sincre} = X except o-sincre; os-X = stance X self-reported by other

Feature Set	SVM (%)	LR (%)
s-flirt	**66.53**	**66.53**
o-flirt	61.75	62.52
o-stances	60.07	61.1
os-stances	49.15	50.44
s-stances	61.64	66.23
o-sincre	*56.58*	*56.58*
o-attrct	**74.77**	**74.77**
o-funny	**65.26**	**63.97**
tm	62.73	61.05
tm, o-sincre	64.24	62.99
tm, o-sincre, o-attract	**75.85**	**75.96**
o-all, tm	73.71	75.86
o-all - {o-sincre}, tm	73.98	75.64
s-flirt, o-attrct, o-funny, tm	74.89	75.68
o-attrct, o-funny, tm	**75.83**	**75.94**

TABLE 2. Correlation between outcomes and **willng**. Bolded are the most relevant other labels.

Feature	Pearson's R	p-value
timemtch	**0.2667**	**1.033e-29**
s-fndly	0.135	7.055e-09
s-flirt	**0.3714**	**8.375e-61**
o-fndly	0.1632	7.421e-12
o-flirt	**0.2553**	**2.761e-27**
o-awk	-0.1488	4.491e-10
o-assert	0.1185	7.197e-07
o-attrct	**0.5298**	**1.756e-126**
o-sincre	0.1899	1.363e-15
o-intell	0.2198	1.797e-20
o-funny	**0.3294**	**2.541e-45**
o-ambits	0.1535	1.225e-10
o-crteos	0.1411	3.358e-09

locutor is perceived. Thus, our conclusion is that willingness to give out e-mail is a statement of attraction, not of trust. We will therefore not focus on this variable **(willng)**.

2) Werewolf

In the Werewolf game [6], players are assigned werewolf and villager (non-werewolf) roles. The werewolves know each other's roles, while the non-werewolves do not know any other player's role. The game proceeds in turns, where each turn has an in-game night and an in-game day. Each night, the werewolves 'eat' one non-werewolf, removing them from the game. The next day, all the remaining players discuss in order to decide who is a werewolf. Then, they vote on whom they believe is a werewolf, and the player with the most votes is removed from the game.

The werewolves win the game as soon as they outnumber the villagers, and the villagers win the game as soon as there are no werewolves left. In general, games start with fewer werewolves than half the number of villagers.

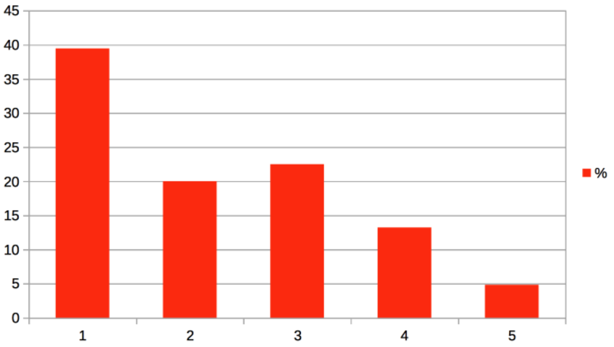

Figure 2. Label distribution for perceived sincerity in Werewolf

The data consists of 86 games with an average game length of 205 utterances. The games were played online, so players only communicated through writing.

We do not have ground truth for perceived sincerity for this dataset. Instead, we use players' votes and vote outcomes as a measure of perceived sincerity.

The more turns a player lasts without being voted out, and the fewer votes that player had in general, the higher the perceived sincerity. Therefore, perceived sincerity would be proportional to the number of turns a player lasts, and inversely proportional to the number of votes a player has throughout. Players who are not voted at all are assigned the highest perceived sincerity.

We are thus trying to predict the following label: $sincere(w) = \frac{N}{1+p}$, where N is the number of turns a player lasted and p is the player's proportion of the total number of votes. The resulting labels are between 0 and 4, with the distribution in Figure 2.

For the experiments in Section 5, we only consider the non-werewolves' perceptions. This is because werewolves already know who has a deceptive role and who has not.

4. Features

We introduce new features for perceived sincerity. The main insight is that qualities of the participant and of the interactional style affect perception of sincerity. Therefore we use new features that we hypothesize are indicative of such qualities. We also consider features used before for different tasks, such as: persuasion [22], deception [4], [7], [18], interactional style identification [10], and stance detection other than deception [20], [21].

1) Sincerity perception features

The presence and duration of silent pauses have been found to be associated with perceived sincerity [9]. In our datasets, we have no means to identify silent pauses. We have, however, considered other features used in this work (Table 3): false starts, repetitions, utterance duration, and

TABLE 3. CURRENT FEATURES OF PERCEIVED SINCERITY [9]

Feature	Description	Comment/Example
changes	no. false starts ('–')	*Are you–I'm a*
repeat	no. repeated words	*Don't you –you*
utterances	number of utterances	
duration	average utterance duration	(SpeedDate only)
sprate	ratio of syllables to duration	(SpeedDate only)

TABLE 4. FEATURES OF PERSUASIVE LANGUAGE [18]

Feature	Description	Example
hedge	no. hedges	*somewhat, nearly, maybe*
ttr	type-token ratio	
hesitations	no. hesitations	*uh, um, er*

speech rate (number of syllables per second). Incomplete words are marked in the SpeedDate dataset, so we used them as a measure of false starts. We also used the start and end time stamps and counted the syllables in each utterance to determine the duration and speech rate.

2) Persuasive language

Our intuition is that gaining trust is to some extent equivalent to persuading a person of one's sincerity.

Language complexity is a marker of persuasive language [22], and one measure of language complexity is the type-token ratio (TTR). The type-token ratio is the ratio between word types and word tokens and gives an indication of how rich a speaker's vocabulary is. On the other hand, hesitations (e.g. *um, uh*) and hedges (e.g. *rather, kind of*) are indicative of weak, unconvincing language (Table 4).

3) Deception and honesty

The research in Linguistic Inquiry and Word Count (LIWC) [19] identifies the following features of deception: words that denote movement (motion words), sensing (sense words), and negative emotions. The authors also identify the following as associated with honesty: exclusion words (e.g. *only, either, without*), self-references, and cognitive complexity. Table 7 shows features of cognitive complexity: conjunctions, prepositions, exclusion words, cognitive words (e.g. *believe, think, recall*) , and long words. Conjunctions are used to integrate between different aspect of a cognitive task, whereas exclusion words are used to differentiate between competing ideas. Psycholinguistic investigations [18] also show a link between deception and utterance length, negative statements, and first-person singular pronoun usage.

Reality monitoring [18] underlines how language is different when recounting a true memory (recalling reality) versus a false one (a mental representation of a fabricated story). People talking about a true memory tend to focus on the attributes of the recalled stimulus (e.g. shape, location, color), whereas people talking about a false memory tend to use more cognitive words and hedges.

Other work [4] has found hesitations correlated with honest rather than deceptive stances. This supports the intuition that deception is a type of persuasive communication, where deceitful speakers use more persuasive language and tend to avoid hesitations. Our previous investigations [23] found motion words, negative emotions, prepositions, and exclusions to be indicative of a deceptive stance in the Werewolf game. These results, showing that deceitful speakers avoid weak language (hedges) and display features of cognitive complexity (prepositions, exclusions), point to deceptive language being more planned and purposely powerful. Positive emotion and pleasantness have also been found indicative of deception [4], [7].

TABLE 5. FEATURES OF DECEPTION [4], [7], [19]

Feature	Description	Example/Comment
motion	no. words denoting motion	*arrive, went, walk*
sense	no. sensing words	*view, saw, hear*
posemo	no. positive emotion words	*sleek, sagely, regal*
negemo	no. negative emotion words	*foul, protest, hate*
avgwords	avg. words per utterance	
neg	number of negations	*no, none, nor*

TABLE 6. FEATURES OF HONESTY [4], [19]

Feature	Description	Example/Comment
exclusion	no. exclusion words	*but, without, either*
self	self-references	*I, mine, myself*
cogcom	cognitive complexity	Table 7
hesitations	no. hesitations	*uh, um, er*

We summarize these features in Tables 5 and 6. We use the MPQA lexicon [24] for sentiment polarity.

4) Interaction style and stance detection

Previous work analyzed stance and perception of interactional style in dating [10], [20], [21]. While we did not find significant correlations between stances and perceived sincerity, sincerity can be seen as similar to a stance, so we consider previously used stance detection features.

We use swear words, sexual words, and words that denote anger, assent, and dissent (Table 8). Swear and dissent words can be seen as markers of speaker confidence, as mentioned in literature [25]: taking an extreme position, whether in agreement or disagreement, signals sincerity. We use LIWC [19] to assign values to those features for each utterance. The SpeedDate corpus has laughter specifically marked, so we use laughter as a feature for this dataset. We use discourse features as well, e.g. the number of turns and number of questions (as question marks).

We use existing word lists [26] for our swear word lexicon. Previous work [27] found that use of profanity mitigates perception of lying and deception. The neurological basis [28] is that, while the brain's language center is in the left

TABLE 7. FEATURES OF COGNITIVE COMPLEXITY [19]

Feature	Description	Example/Comment
cog	words denoting cognition	*think, plan, believe*
exclusion	no. exclusion words	*but, without, either*
conj	no. conjunctions	and, but, whereas
prep	no. prepositions	to, with, above

TABLE 8. STANCE/INTERACTIONAL STYLE FEATURES [10], [20], [21]

Feature	Description	Example/Comment
swear	no. swear words	*heck, damn, crap*
anger	no. anger words	*stupid, sucks, hate*
assent	no. assent words	*yes, cool, agree*
politeness	no. polite words	*please, thank, excuse*
dissent	no. dissent words	*no, actually, yes but*
laugh	laughter instances	(SpeedDate only)
you	no. references to the other	*you'd, you'll, your*
us	no. us words	*our, we, ourselves*
insight	no. insight words	*think, feel, figure*
qmark	number of question marks	
turns	number of turns	

hemisphere of the cerebral cortex, swearing is controlled by the limbic system, responsible of processing emotions. So swearing is processed in the emotion center, creating the impression that it isn't planned or scripted, but rather more authentic and passionate. Other explanations [29] note that swearing carries a social risk, and that sharing unharmful but socially unacceptable behavior is a catalyst of bonding.

5) New features

Inspired by the observation that pleasantness can be indicative of deception [4], [7], we add **politeness** as a marker of pleasantness. Politeness is a guiding principle in communication according to which people try to show respect towards others (positive politeness) and minimize their impositions on others (negative politeness) [30]. Elements of positive politeness are gratitude (e.g. *thank you*), positive sentiment (e.g. *wonderful*), solidarity, and inclusiveness. Elements of negative politeness are indirection (e.g. *by the way, could you possibly*), modalities (e.g. *would, could*), apologies (e.g. *sorry, excuse*), and other politeness markers such as *please*. We use the Stanford politeness system [30] to calculate the odds of politeness for each utterance.

Many of the features in Table 8 are also associated with pleasantness (e.g. laughter, anger, dissent). The use of *us* words (e.g. *our, we*), as well as showing interest in **social** issues, can indicate solidarity and inclusivity. Thus, we use social words (e.g. *mate, talk*) (as assessed by LIWC [19]) as another feature of perceived sincerity (Table 9).

Regarding assent, previous work [25] has found that agreement influences perceived competence rather than sincerity, and that the more extreme a position one has on an issue, the more sincere they are perceived. We use **intensifiers** (e.g. *very, super, highly*) and **exclamations** as markers of how extreme one's position is. On the opposite side, we included features indicative of uncertainty, such as tentative (e.g. *maybe, perhaps, guess*), anxiety (e.g. *worried, nervous*), and discrepancy words (e.g. *should, would, could*). We used LIWC [19] to assess use of such words. We included causation words, as well as commas.

As observed in [31], many of these features are also markers of **openness**. In an intuitive sense, we expect openness to be related to perceived sincerity. While openness is helpful in establishing trust, too much self-disclosure is perceived as unpleasant by the listener [32]. On the other hand, engaging in self-disclosure is perceived as a sign of a positive impression on the part of the speaker [32]. As additional markers of openness, we include (Table 9) words indicating **informal** speech, **affective** processes (e.g. *happy, cried*), and **family** (e.g. *mother, sister*). Informal speech is reflected in: assent words (e.g. *yes, cool, agree*, see Table 8), fillers (e.g. *blah, meh, you know*), netspeak (e.g. *dat, tha, tho, liek*), and swear words (e.g. *heck, damn, crap*, see Table 8). We use LIWC [19] for those features.

TABLE 9. NEW FEATURES

Feature	Description	Example
politeness	no. polite words	*please, thank*
social	no. words indicating social concerns	*mate, talk, child*
intens	no. intensifiers	*very, super, highly*
wow	no. exclamation marks	
anxiety	anxiety words	*worried, nervous*
discrepancy	discrepancy words	*should, would, could*
tentative	tentative words	*maybe, perhaps, guess*
affect	no. words indicating affective processes	*happy, cried*
informal	no. words indicating informal speech	assent, fillers netspeak, swear,
family	no. family-related words	*mother, sister*
ing	words that end in *-ing*	
shehe	no. third-person singular pronouns	*she, her, him*
period	no. periods	
focuspast	verbs in past tense	*went, ran, had*
focuspresent	verbs in present tense	*goes, is, has*
focusfuture	verbs in future tense	*will, gonna*
comma	no. commas	
cause	causation words	*because, effect, hence*

On affective processes, we want to underline the distinction from emotion, which we also use as a feature (Section 5). Affect is the conscious subjective experience, while emotional affect is the unconscious component [33]. Features of emotion usually refer to the unintentional display of emotion. Features of affect refer to intentionally discussing the subjective experience, which is a sign of openness.

We considered words ending in **-ing** as marker of vagueness, which may be due to deflection. Words ending in -ing are also among the top tf-idf scoring words. Another sign of deflection can be referring to others who are not present. We therefore counted third-person singular pronouns.

Other features we considered were simple periods ('.'), as a measure of how emotional one's self-expression is, and whether the speaker's focus is on past, present, or future.

5. Experimental results

We will first discuss willingness to give contact information in the SpeedDate corpus and the connection with perceived sincerity. Then we show experimental results for perceived sincerity in the SpeedDate and Werewolf datasets.

1) Perceived sincerity in Speed Dating As a start, we want to see how well a bag-of-words (BOW) baseline classi-

TABLE 10. 10-FOLD CROSS VALIDATION ACCURACY OF BINARY CLASSIFICATION OF **o-sincre** WITH PSYCHOLINGUISTIC FEATURES. T=3 IS THE CUTOFF THRESHOLD FOR LABEL BINARIZATION: [1–3] NEGATIVE, [4–10] POSITIVE (55.73% NEGATIVE). BOLDED ARE THE BEST RESULTS.

Feat.	NB (%)	SVM (%)	LR (%)
BOW	44.94 ± 3.67	50.03 ± 5.12	43.93 ± 6.13
All	49.43 ± 4.94	53.98 ± 0.11	49.94 ± 6.38
SD1	53.92 ± 0.19	**53.98 ± 0.1**	**52.77 ± 4.4**
SD2	**54.33 ± 1.4**	**53.98 ± 0.1**	**52.77 ± 4.7**
SD	53.98 ± 4.13	53.98 ± 0.11	49.49 ± 3.83

TABLE 11. 10-FOLD CROSS VALIDATION F1, RECALL, AND PRECISION OF BINARY CLASSIFICATION OF **o-sincre** (T=3). FOR EACH FEATURE SET, WE SHOW THE BEST RESULT ACROSS ALL CLASSIFIERS (NB, LR, SVM). PSI – ALL PSYCHOLINGUISTIC FEATURES. BOLDED ARE THE BEST RESULTS.

Features	F1	Precision	Recall
BOW	39.97 ± 7.84	47.64 ± 17.58	40.63 ± 13.61
Psi	**54.8 ± 6.78**	46.43 ± 4.27	**67.9 ± 13.71**

fier performs. We use Naive Bayes (NB), SVM, and logistic regression (LR) as classifiers, with 10-fold cross-validation. We report the results for binary classification in Table 10 and 11 and for multi-class classification in Table 12. For the binary classification problem, we use binarization threshold T=3, as it results in the most balanced label distribution.

We repeat the experiments using the psycholinguistic features in Section 4. We normalize all features.

We perform feature selection and show the results of the best performing sets, which we call SD1 and SD2, in Table 10. All psycholinguistic feature sets significantly outperform the bag-of-words baseline. SD1 consists of average number of words per utterance and long words, positive and negative emotions, self-references, as well as TTR. SD2 adds insight words, swear words, dissent words, and number of turns.

By performing correlation analysis, we discover significant correlations between sincerity perception and: type-token ratio, period use, positive emotion, focus on the past, affect, emotion, use of long words, informal words, hesitations, social words, exclamations, sense words, and use of third-person singular. We call this feature set SD.

In addition to accuracy (Table 10), we also show precision, recall, and F1 (Table 11). We can see that while BOW performs slightly better in precision, the psycholinguistic features have a much better recall and F1.

Since our gold standard is *perceived* sincerity, a false negative is a situation where the listener perceives the speaker as sincere, but our system thinks he perceives him as insincere. False negatives put the listener at a higher risk of deception; e.g., if the speaker was indeed deceptive, and the system thinks the listener is already wary, it may not issue a warning. On the other hand, if the listener is distrustful, but the system labels him as trusting (a false positive), then the worst case scenario is an unnecessary warning. Therefore, false negatives are more costly than false positives, so recall is a better measure for decoding perceived sincerity.

TABLE 12. 10-FOLD CROSS VALIDATION ACCURACY OF MULTI-CLASS CLASSIFICATION OF **o-sincre** WITH PSYCHOLINGUISTIC FEATURES. THE MAJORITY BASELINE IS 20.99%. ALL – ALL PSYCHOLINGUISTIC FEATURES. WE BOLDED THE BEST RESULTS.

Features	NB (%)	SVM (%)
BOW	13.92 ± 2.77	17.97 ± 2.64
All	15.47 ± 3.16	20.94 ± 0.23
SD	**18.12 ± 3.14**	**20.94 ± 0.23**

We also show experiments for the multi-class classification problem (Table 12). We show experimental results for bag-of-words features, all psycholinguistic features, and features correlated to the label. Again, the psycholinguistic features outperform the bag-of-words features.

Overall, psycholinguistic features perform better than the bag-of-words baseline. The reason is that the feature space is much sparser, and the new features capture connections between language and psychological processes. We will discuss these connections in detail in Section 6.

Feature correlations per gender. The following features were significant for the SpeedDate dataset: number of periods, hesitation and disfluencies, emotion, positive emotion, affective, informal, social, sense, and long words, focus on the past, exclamations, third-person singular pronouns (negative); and type-token ratio and politeness (positive).

We find that women and men rely on different cues to judge the other's sincerity. For men, exclamations and focus on the past, others, and family are seen as insincere. These are markers of openness, which may be seen as untimely.

On the other hand, women rely on many more cues to make a decision. In addition to the SD features above, the following were negatively correlated with perceived sincerity: verbosity (duration, number of utterances, number of turns), cognition (insight and cognitive words), conflict words (anxiety and dissent words), laughter, focus on the present, weak language (tentative words, discrepancies), and discourse (comma use, conjunctions, causation words). Exclamations and focus on the past were not relevant, while politeness was negatively correlated to perceived sincerity.

Some of these features are understandable in the context of deception cues. Verbosity and complex discourse can be seen as deflective, while weak language is a marker of uncertainty, which makes people distrustful. Since they are more relevant for women than in general, we can conclude that women may be more alert to possible deception.

With the observation that talking about one's past and future require more openness, one conclusion is that women require more openness to judge the speaker as sincere.

On the other hand, while a marker of authenticity, conflict diminishes pleasantness. Women may see positive language as more sincere. This can pose a risk in online interactions, as pleasantness is also associated with deception.

2) Perceived sincerity in the Werewolf game. We perform multi-class classification, with the label distribution in Figure 2, as well as binary classification, with threshold T=2 (59.46% negative).

TABLE 13. 10-FOLD CROSS VALIDATION ACCURACY OF MULTI-CLASS CLASSIFICATION. BOLDED ARE THE BEST RESULTS

Features	NB (%)	SVM (%)
BOW	35.73 ± 5.55	36.44 ± 5.39
All	32.16 ± 2.3	35.73 ± 4.96
SD1	28.44 ± 8.23	35.21 ± 7.48
W1	38.57 ± 1.7	36.5 ± 6.4
W2	35.22 ± 7.66	39.34 ± 8.48
SD2	$\mathbf{35.4 \pm 6.01}$	$\mathbf{38.8 \pm 8.32}$
SD	37.35 ± 4.06	36.15 ± 7.18
WW	$\mathbf{38.61 \pm 3.91}$	$\mathbf{38.81 \pm 4.65}$

TABLE 14. 10-FOLD CROSS VALIDATION ACCURACY OF BINARY CLASSIFICATION. BOLDED ARE THE BEST RESULTS.

Features	NB (%)	LR (%)
BOW	57.83 ± 8.88	56.77 ± 6.29
All	57.61 ± 7.83	62.1 ± 8.51
SD	58.18 ± 7.31	$\mathbf{59.81 \pm 3.86}$
WW	$\mathbf{59.28 \pm 6.44}$	58.23 ± 5.5

We only found significant correlations with the assigned perceived sincerity label for hesitations and swear words (positive correlation). We call this feature set WW.

We perform feature selection to identify the best performing feature subset. One such subset, which we call W1, consists of cognitive words, positive and negative emotions, self-references, long words, and TTR. Another subset, which we call W2, adds the following features to W1: *us* words (see Table 8); insight, dissent, sexual, and swear words; laughter, hesitations, questions, and turns.

Table 13 summarizes the best results. We also show results for the best performing feature subsets for the dating dataset (SD1, SD2) and for the features correlated with perceived sincerity in both datasets.

Several subsets of psycholinguistic features outperform the bag-of-words baseline. The correlated features for both Werewolf and SpeedDate datasets perform best.

6. Discussion

To answer Question 1 of Section 1, our experiments show that perceived sincerity in dialogues, using only written language as input, is a hard problem. Shallow classifiers using both bag-of-words and psycholinguistic features perform poorly. The gap between the classifiers' performance and the gold standard might be explained by the lack of non-verbal features and their combination with verbal ones, which would support the 7% Rule.

Since context and complex cognitive processes are involved in deception and sincerity, it may be useful to explore the role of deeper interactions not explicitly observed or labeled, for example by adding latent variables to the model.

In the following, we will answer Question 1.

For speed dating, a rich vocabulary (type-token ratio) and politeness were positively correlated with the label. These are markers of persuasive language [22]. This suggests that powerful language is perceived as more sincere.

Inspecting the odds ratios of logistic regression: use of self-references, insight, dissent, and swear words were found indicative of perceived sincerity. Since self-references are indicative of honesty, it is not surprising that the listener picked up on such cues to decide the speaker was honest. As expected, insight, swear words, and dissent indicate perceived sincerity.

Some of the features indicative of perceived insincerity were related to verbosity (number of turns, long words), cognition, uncertainty (number of periods, hesitations, and disfluencies). This is not surprising, since verbosity (e.g. average word count) is indicative of deception, and weak, uncertain language does not inspire trust.

On uncertain language, previous research [4] has shown that, while filled pauses signal discomfort with a topic and are thought of as markers of deception, they correlated more strongly with truthful statements. The fact that people are more likely to label these statements as insincere shows that humans may not be good at perceiving sincerity.

Other markers of perceived insincerity were related to pleasantness (politeness, positive emotion, affective and social words) and openness (emotion, informal, exclamations). It is interesting to note that pleasantness can indeed be indicative of deception [4], [7]. Also, it is possible that, given the context (4 minute conversations), too much openness can be perceived as unwarranted and therefore, inauthentic.

Other features can be described as story-telling features (focus on the past, third-person singular pronouns), as they are used to relate events about other people. While relating events from the past is conducive to bonding, focusing on other people is a form of deflection from oneself. The combination can be seen as a one-sided request for trust.

For the Werewolf game, verbosity plays less of a role, and use of cognitive words was more important, a feature indicative of honesty. In addition, features of pleasantness turned out to be more important as well.

Features indicative of perceived sincerity are positive emotion, number of turns, hesitations, and number of swear words. Features indicative of perceived dishonesty were dissent, cognitive words, and insight words. These findings are similar to the case of the SpeedDate corpus, with the exception of dissent, insight words, and hesitations.

Hesitations are, as noted before, markers of honesty, and it is possible that players of Werewolf are more alert to cues of deception and honesty. For dissent, one reason why it is seen as insincere may be that, in the Werewolf game, the players who talk more (usually the werewolves themselves) are also those who take charge of the problem of werewolf identification and are thus more likely to use insight words. As for disagreement, in this setting, disagreement is equivalent to defending oneself or other players, often in the face of group consensus, which can be seen as suspicious.

It is also important to note that, for dating, the participants do have access to non-verbal cues to rely on, even if the transcriptions do not contain them. They also have to pay attention to much more stimuli and have less time to process them. On the other hand, in the Werewolf game, the interaction between players is only through written text. It is therefore to be expected that Werewolf players rely more on written language cues than the people in the SpeedDate dialogues, and have more time to give them more weight.

7. Conclusion

We draw attention to the problem of perceived sincerity detection in written dialogues, which has received very little attention. The low classification accuracy for both baseline bag-of-word and psycholinguistic features show that this is a difficult problem that requires further investigation.

We find that several psycholinguistic features are similarly correlated to perceived sincerity across domains, such as language complexity (TTR), cognitive processing (cognitive words, long words), strong opinion (dissent, swear words), positive emotion, and verbosity. Deception cues help in identifying perceived sincerity only to the extent that they are not overridden by pleasantness features.

The classification performance may be improved by using sequential models of conversation and more explicit modeling of cognitive state – for example, sincerity, openness, cognitive complexity, and the strength of opinion can be modeled as latent variables. It may also be useful to model per-person priors on these variables.

Another direction for future work is understanding the differences in the perception of sincerity across several other categories, such as age and education. The dialogue participants in our datasets were in the same range with respect to both, therefore a related direction of future work is collecting dialogue data where participants display a higher diversity across these new categories.

References

[1] A. Mehrabian, *Silent Messages: Implicit Communication of Emotions and Attitudes*.

[2] S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab, "Searching for safety online: Managing" trolling" in a feminist forum," *The Information Society*, vol. 18, no. 5, pp. 371–384, 2002.

[3] C. Hardaker, "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions," 2010.

[4] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis *et al.*, "Distinguishing deceptive from non-deceptive speech." in *Interspeech*, 2005, pp. 1833–1836.

[5] F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke, "Detecting deception using critical segments." in *Interspeech*, 2007, pp. 2281–2284.

[6] H. Hung and G. Chittaranjan, "The wolf corpus: Exploring group behaviour in a competitive role-playing game," in *ACM Multimedia*, 10 2010.

[7] V. Niculae, S. Kumar, J. Boyd-Graber, and C. Danescu-Niculescu-Mizil, "Linguistic harbingers of betrayal: A case study on an online strategy game," in *ACL*, 2015.

[8] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of Interspeech*, 2016.

[9] J. Loy, H. Rohde, and M. Corley, "Lying, in a manner of speaking," in *Proceedings of speech prosody*, vol. 8, 2016.

[10] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," in *NAACL '09*.

[11] C. L. Toma and J. T. Hancock, "What lies beneath: The linguistic traces of deception in online dating profiles," *Journal of Communication*, vol. 62, no. 1, pp. 78–97, 2012.

[12] R. E. Guadagno, B. M. Okdie, and S. A. Kruse, "Dating deception: Gender, online dating, and exaggerated self-presentation," *Computers in Human Behavior*, vol. 28, no. 2, pp. 642–647, 2012.

[13] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *ACL-IJCNLP 2009*, pp. 309–312.

[14] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam." in *HLT-NAACL*, 2013, pp. 497–501.

[15] C. L. Toma and J. T. Hancock, "Reading between the lines: linguistic cues to deception in online dating profiles." in *CSCW*, K. I. Quinn, C. Gutwin, and J. C. Tang, Eds. ACM, 2010, pp. 5–8.

[16] E. Fitzpatrick, J. Bachenko, and T. Fornaciari, "Automatic detection of verbal deception," *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 3, pp. 1–119, 2015.

[17] H. Kaya and A. A. Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," *Interspeech 2016*.

[18] S. Porter and J. C. Yuille, "The language of deceit: An investigation of the verbal clues to deception in the interrogation context," *Law and Human Behavior*, vol. 20, no. 4, pp. 443–458, 1996.

[19] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[20] R. Ranganath, D. Jurafsky, and D. McFarland, "It's not you, it's me: detecting flirting and its misperception in speed-dates," in *EMNLP 2009*.

[21] R. Ranganath, D. Jurafsky, and D. A. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Computer Speech & Language*, vol. 27, no. 1, pp. 89–115, 2013.

[22] L. A. Hosman, "Language and persuasion," in *The Persuasion Handbook: Developments in Theory and Practice*, J. P. Dillard and M. Pfau, Eds. Sage Publications, 2002.

[23] C. Gîrlea, R. Girju, and E. Amir, "Psycholinguistic features for deceptive role detection in werewolf," in *NAACL HLT 2016*.

[24] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *HLT '05*.

[25] R. Eisinger and J. Mills, "Perception of the sincerity and competence of a communicator as a function of the extremity of his position," *Journal of Experimental Social Psychology*, vol. 4, no. 2, pp. 224–232, 1968.

[26] R. Jones, "noswearing.com," http://www.noswearing.com/dictionary, 2012, [Online; accessed 30-April-2017].

[27] G. Feldman, H. Lian, M. Kosinski, and D. Stillwell, "Frankly, we do give a damn: The relationship between profanity and honesty," *Social Psychological and Personality Science*, 2017.

[28] B. Bergen, *What the F: What Swearing Reveals About Our Language, Our Brains, and Ourselves*. Basic Books, 2016.

[29] M. Adams, *In Praise of Profanity*. Oxford University Press, 2016.

[30] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, "A computational approach to politeness with application to social factors," in *ACL*, 2013.

[31] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[32] A. P. Bochner, *On the efficacy of openness in close relationships*. Transaction Books New Brunswick, NJ, 1982.

[33] J. Panksepp, "Affective consciousness: Core emotional feelings in animals and humans," *Consciousness and cognition*, vol. 14, no. 1, pp. 30–80, 2005.