

Detecting Bot-Generated Text by Characterizing Linguistic Accommodation in Human-Bot Interactions

Paras Bhatt and Anthony Rios

Department of Information Systems and Cyber Security

University of Texas at San Antonio

{Paras.Bhatt, Anthony.Rios}@utsa.edu

Abstract

Language generation models' democratization benefits many domains, from answering health-related questions to enhancing education by providing AI-driven tutoring services. However, language generation models' democratization also makes it easier to generate human-like text at-scale for nefarious activities, from spreading misinformation to targeting specific groups with hate speech. Thus, it is essential to understand how people interact with bots and develop methods to detect bot-generated text. This paper shows that bot-generated text detection methods are more robust across datasets and models if we use information about how people respond to it rather than using the bot's text directly. We also analyze linguistic alignment, providing insight into differences between human-human and human-bot conversations.

1 Introduction

Bots are useful in a wide variety of applications areas including business (Kaczorowska-Spychalska, 2019), education (Kerlyl et al., 2006), and health (Yadav et al., 2019; Liednikova et al., 2020). For instance, Yadav et al. (2019) studied the use of chatbots as a drop-in first-point-of-contact for women in India seeking breastfeeding information. Similarly, researchers have studied bots to answer COVID-19-related questions (Oniani and Wang, 2020) and screen individuals for risks of contracting the virus (Martin et al., 2020). Overall, the wide availability of software packages, tools, and pre-trained models has democratized the creation of bots.

Even with the increasing interest in bots for social good (e.g., COVID-related chatbots), there is still a concern regarding their abuse to spread misinformation, be used for targeted discrimination, deceive users, and perform fraud (Daniel et al.,

2019). Given the potential good and harm bots can create, it is essential to study how the bots *should* act and how people *do* interact with such bots for specific applications.

Automatically identifying bots online is well studied (Garcia-Silva et al., 2019; Herzig et al., 2019; Kosmajac and Keselj, 2019; Ippolito et al., 2020; Jawahar et al., 2020). Bot detection methods rely on two forms of information (Orabi et al., 2020): behavior and content. Behavior relates to measuring how often bots post, the time posts are created, and conversational network structures (Beskow and Carley, 2018). Content involves using the bot's text directly. Focusing on text suffers from generalization issues, making it challenging to detect bots that discuss different topics or operate in different domains. Likewise, behavioral approaches assume that bots will behave differently than people at a superficial level, such as posting more often than humans. Even in the network analysis of conversation structures (Beskow and Carley, 2018), many of the human-bot interactions on social media are not human-like interactions. For instance, it is understandable that humans will interact with a bot that converts pounds to kilograms differently than other humans. Given bots' current use-cases ranging from counselors to healthcare information providers, it is vital to understand how humans and bots interact beyond trivial applications.

Before looking at how humans interact with bots, it is essential to understand how humans interact. This paper focuses on Communication Accommodation Theory (CAT) and general language use to analyze interactions. CAT is used to study language use in various domains to understand human behavior (Giles et al., 1973; Tausczik and Pennebaker, 2010). More specifically, we study linguistic accommodation—where speakers come to talk more (or less) similarly as they interact—

which has been analyzed in mental health support on Reddit, showing a positive link with informational and emotional support (Sharma and De Choudhury, 2018). The alignment of two people in terms of linguistic style has positively predicted successful outcomes of negotiations (Taylor and Thomas, 2008), and doctors are recommended to accommodate perspective to improve patient care (Kline and Ceropski, 1984; Wood, 2019). The way people use and accommodate pronouns can indicate power, where high-status individuals use “I” less and “you/we” words more than low-status individuals (Kacewicz et al., 2014).

Linguistic accommodation in human-bot interactions has been studied in system design, showcasing that there is a strong link between user experience and language style of the system (Chaves et al., 2019; Chaves, 2020; Thomas et al., 2020). Recent research has also shown a link between language style in chatbots and user engagement for e-commerce (Elsholz et al., 2019). Our focus is to understand real long open-domain human-bot interactions better. Our study can also influence how bots are detected on social media and provide a better understanding of how human-human interactions differ from human-bot interactions is essential for bot development. For instance, when should bots accommodate towards the user? Can we tell how well a system performs for a specific task based on whether the user accommodates the bot? If human-bot conversations are linguistically different than human-human interactions, research from human-human studies may not generalize to human-bot interactions.

Toward addressing the potential societal impacts of open-domain bots, this paper addresses the following research questions (RQs) in the context of bot Detection:

- RQ1.** How do humans and bots align in human-bot interactions?
- RQ2.** How does the alignment in human-bot interactions compare to the alignment in human-human interactions?
- RQ3.** Are differences in alignment between human-human and human-bot interactions similar across domains and language generation learning methods?

Overall, we analyze whether human users’ language changes stylistically with bots as compared

to human-human conversations. Furthermore, we show that small differences in language style provides robust information, compared to the bots’ language patterns, to accurately detect bots.

2 Related Work

Bot Detection. Detecting bots in the wild is a widely studied problem (Garcia-Silva et al., 2019; Herzig et al., 2019; Kosmajac and Keselj, 2019; Ippolito et al., 2020; Jawahar et al., 2020). Bot detection methods rely on two forms of information (Orabi et al., 2020): Content and Behavior. Garcia-Silva et al. (2019) studied how pre-trained language models perform for the task of bot detection. Kosmajac and Keselj (2019) developed “language-independent” stylistic features that measure language diversity to detect bots. Knauth (2019) empirically explored content and behavioral features for bot detection. Beskow and Carley (2018) analyzed human-bot interactions by developing methods to detect bots using network analysis of conversation structures. Much of the prior work has focused on detecting bots “in the wild.” Thus, many of the interactions between bots and humans are superficial (e.g., receiving movie quotes from a bot). This paper differs from prior work in two ways. First, rather than using content from the bot or general behavioral information (e.g., post frequency), we try to detect bots by analyzing how humans respond to them (compared to human-human interactions). Second, rather than exploring bots in the wild, we explore two types of datasets: one where researchers instruct participants to interact with bots and humans in the same way, and another dataset where researchers instruct participants to converse with a bot.

Analyzing Human-Human Conversations. Linguistic alignment is known to interact with a wide array of social factors. For instance, the level of alignment of people’s linguistic style in a conversation has been claimed to be affected by their relative social power (Gnisci, 2005; Xu et al., 2018; Danescu-Niculescu-Mizil et al., 2011). Cooperative decision-making tasks are positively related to the participants’ linguistic convergence (Fusaroli et al., 2012; Kacewicz et al., 2014). Recently, Sharma and De Choudhury (2018) analyzed mental health support forums on Reddit, showing that linguistic accommodation is positively linked with informational and emotional support. Similarly,

Taylor and Thomas (2008) analyzed negotiation outcomes and linguistic alignment, showing that alignment positively predicted successful negotiation outcomes. There is also the support of linguistic alignment by doctors to enhance patient care by improving trust and adherence to a treatment plan (Kline and Ceropski, 1984; Wood, 2019). Alignment is a powerful tool that can be used to improve understanding, trust, and potentially patient outcomes. Therefore, understanding *when* and *how* people should align with each other is an important area of research. Furthermore, it is essential to draw a line between engaging with a culture and language style and appropriating it, thereby potentially causing harm rather than building rapport with the partners in a conversation.

With the goal of understanding human-human interactions, there has also been progress in linguistic alignment measurement techniques (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011; Jones et al., 2014; Wang et al., 2014; Doyle and Frank, 2016; Shin and Doyle, 2018). Danescu-Niculescu-Mizil et al. (2011) presented an easy-to-compute expression that measures the increase in the conditional probability given that a conversational partner has used it. An issue with the method proposed by Danescu-Niculescu-Mizil et al. (2011) is that it assumes messages between two people have similar length. To overcome this limitation, Doyle and Frank (2016) introduced the Word-Based Hierarchical Alignment Model (WHAM), a hierarchical graphical model where the parameters are learned using Bayesian inference. Another method called the Simplified Word-Based Alignment Model (SWAM) was recently proposed by Shin and Doyle (2018). SWAM attempts to compare alignment between different groups when the alignment scores are assumed to differ substantially. Unfortunately, SWAM only estimates group-level alignment, not conversation-level between two specific users/bots. Linguistic alignment has also been shown to be predictive of specific tasks. In this paper, we use the method proposed by Danescu-Niculescu-Mizil et al. (2011) to estimate linguistic alignment. For instance, Niven and Kao (2019) use alignment features to predict discourse acts.

Analyzing Human-Bot Conversations. Research studying the interaction between bots and humans has been explored from a wide array of

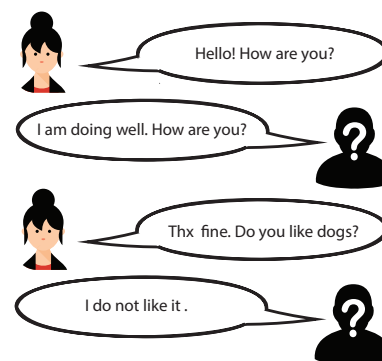


Figure 1: This figure depicts an potential conversation between a known human user and an unknown user. The unknown user may be a human or a bot.

perspectives. For example, systems that use emotionally expressive interjections (“wow”, “ahem”) in their text to speech responses can significantly improve the user experience (Cohn et al., 2019). Given the popularity of bots in application areas from business (Kaczorowska-Spychalska, 2019) to healthcare (Pieraccini et al., 2009), it is also important to understand how language generation style and alignment impacts their intended use. There has been a recent interest in analyzing accommodation and similar concepts in human-bot interactions. For instance, Ahn et al. (2020) show that humans will match *code-switching* patterns introduced by a chat system. Moreover, users tend to have positive reactions towards systems that code-switch. Ma and Lalor (2020) measure *lexical entrainment* between a specific Reddit bot and users. They find that sentiment of bot has a positive effect on the sentiment of the humans response. Furthermore, Ma and Lalor (2020) show that human responses tend to overlap with the bots original post. Compared to prior work studying accommodation-related aspects of human-bot interactions, this paper differs in two ways. First, just analyzing linguistic alignment, we ground our study of human-bot interactions in the real-world task of bot detection. Second, we analyze cross-domain generalization of accommodation patterns in human-bot interactions, where cross-domain includes types of bots (e.g., retrieval and transformer-based models), data collection procedures (e.g., sampling bias), and conversation topics.

3 Datasets

An overview of the task we introduce in this paper is shown in Figure 1. We formulate a bot de-

Dataset Name	# D	# u	Avg. u	Avg. W
ConvAI2 Data				
INTERMEDIATE	291	4317	14.83	6.66
TOLOKERS	3127	39155	12.52	7.05
VOLUNTEER	1111	14623	13.16	6.60
PERSONA-CHAT	18878	278478	14.75	1.85
Control Data				
IRIS	163	5687	34.89	5.46
TICKTOCK	206	5462	26.51	7.07
DailyDialog	13118	102980	7.85	13.08

Table 1: Summary of each dataset, including the number of dialogues (# D), average number of utterances per dialogue (Avg. u), average number of words per utterance (Avg. W) and total number of utterances (U)

tection task between two entities, either a human and bot or a human and human. We assume that one entity is always human, and the other entity is unknown—either a human or a bot. Formally, let $D = [u_1^h, u_2^o, \dots, u_{N-1}^h, u_N^o]$ represents a sequence of utterances, where u_i^h represents the i -th utterance in a conversation and that it was made by a human h . Likewise, u_i^o represents an unknown entity’s utterance (human or bot). N is the total number of utterances in the conversation. Our goal is to develop a classifier $f(D)$ that maps to a class in the set $T = \{\text{human-human, human-bot}\}$, where the human-bot means a bot is a part of the conversation. We formulate two bot detection datasets for this task consisting of three known bot datasets: ConvAI2, WOCHAT, and DailyDialog. The basic statistics of each dataset are shown in Table 1. WOCHAT and DailyDialog are used to form a Control dataset. We describe each dataset below:

ConvAI2. We use four datasets from the Second Conversational Intelligence Challenge (ConvAI2). The motivation behind the ConvAI2 challenge was to develop new approaches towards open-domain chatbots (Zhang et al., 2018; Dinan et al., 2019). Overall, we use two types of ConvAI2 datasets: training datasets containing human-human interactions and evaluation phase datasets containing human-bot interactions. First, we use the PERSONA-CHAT dataset, a collection of human-human interactions where researchers instructed Mechanical Turk users to converse with one another, assuming specific personas (i.e., profile descriptions). The ConvAI2 competition used automatic (e.g., Perplexity) and human evaluation

procedures. The human evaluation procedures involve either paid workers or volunteers that interact with models built using PERSONA-CHAT, after which the volunteers/workers are asked to rate their interactions. Moreover, the human evaluator is instructed to interact with the bots in the same way as the Mechanical Turk users who participated in creating the PERSONA-CHAT dataset.

We use three “evaluation-phase” datasets¹: TOLOKERS, VOLUNTEERS, and INTERMEDIATE. The TOLOKERS used solicited workers to chat with the models, similar to the PERSONA-CHAT’s Mechanical Turk setup. Specifically, TOLOKERS consists of data collected during DeepHack.Chat² hackathon via paid workers using the Yandex.Toloka service. The INTERMEDIATE dataset consists of more dialogues by the bots from DeepHack.Chat, but the interactions come from volunteers. Finally, the VOLUNTEER dataset was collected during the final “wild evaluation” round of the ConvAI2 competition. Human-bot interactions were collected from volunteers through the Facebook Messenger and Telegram APIs.

Control Dataset. One of the research questions we explore in this paper is related to out-of-domain performance. Specifically, can we detect bots based on human response when data were collected under different settings and where the conversation topics differ? To address the out-of-domain research question, we use two additional datasets collected in the The Workshop on Chatbots and Conversational Agent Technologies (WOCHAT) (Kong-Vega et al., 2019). As part of a shared task, the workshop makes several bots available, has participants contribute new bots, and participants interact with the bots providing utterance-level feedback regarding their performance. We use two human-bot interaction datasets released by the organizers: IRIS and the TickTock dataset.³ both IRIS (Banchs and Li, 2012) and TickTock (Yu et al., 2015) are retrieval-based bot variants that were “trained” on different datasets. Unlike the ConvAI2 task, humans are not instructed to converse with the bots as if they are another human. Thus, sometimes humans will ask things such as, “Are you a Robot?”. This point

¹<http://convai.io/data/>

²<http://deephack.me/chat>

³<http://workshop.colips.org/wochat/data/index.html>

provides a unique aspect that increases cross-data differences.

Unfortunately, human-human conversations were not released from the WOCHAT shared task. The focus was annotating human-bot interactions. Thus, we need to augment WOCHAT with real human-human conversations. Our study uses the DailyDialog dataset as our source of human-human interactions (Li et al., 2017). Unlike the ConvAI2-related data, DailyDialog consists of conversations between English learners practicing everyday English dialog in daily life.

Data Processing. We split the data into three groups: Unpaid (U), Paid (P), and Control (C). The U dataset consists of PERSONA-CHAT, INTERMEDIATE, and VOLUNTEER ConvAI2 datasets. The P dataset consists of PERSONA-CHAT and TOLOKERS—the name Paid comes from the fact that the TOLOKERS were paid as part of a crowdsourcing task. We split this group from the others in case this affects conversation behavior. The C dataset consists of IRIS, TICK-TOCK, and DailyDialog. Each dataset is divided into 70%, 10%, and 20% training, validation, and test splits, respectively. It is important to note that the same PERSONA-CHAT training, validation, and test examples are used in both the U and P datasets.

4 Method

To detect human-bot conversations, we explore two types of features below: Content and Stylistic features.

4.1 Content Features

We define content information as features describing “*what*” humans and bots say in their interactions. Specifically, we describe two sets of content features: bag-of-words and embeddings.

Bag-of-words. As a simple baseline, we use TF-IDF-weighted unigrams from a dialog to detect whether a bot is part of the conversation. We explore three settings for the TF-IDF features: human-only, unknown-only, and human-unknown. The *unknown* user in the conversation can be either a bot or a human. Thus, in the human-Only setting, only the human’s unigrams from each dialogue $D_h = [u_1^h, u_3^h, \dots, u_{N-1}^h]$ are used to detect human-bot interactions. The unknown-only setting uses

only the unknown user’s unigrams from each dialogue $D_o = [u_2^o, u_4^o, \dots, u_N^o]$, and the human-unknown setting uses both the human’s and unknown user’s unigrams from each dialogue $D = [u_1^h, u_2^o, \dots, u_{N-1}^h, u_N^o]$.

Embedding Features. We encoded each dialogue D using BERT (Devlin et al., 2019)⁴ by feeding the first 510 WordPieces and then averaging the word representations extracted from the second-to-last layer. We explore three variants, human-Only, Unknown-Only, and human-Unknown. Depending on the variant, the exact WordPieces passed to BERT change. For instance, in the human-Only setting, the first 510 WordPieces made by the human from a concatenation of all of the humans utterances $D_h = [u_1^h, u_3^h, \dots, u_{N-1}^h]$ are passed to BERT. Similarly, all of the unknown user’s utterances are used for the unknown-only variant to obtain $D_o = [u_2^o, u_4^o, \dots, u_N^o]$. In the human-unknown setting, all human and unknown utterances are concatenated in $D = [u_1^h, u_2^o, \dots, u_{N-1}^h, u_N^o]$ before generating features with BERT.

4.2 Stylistic Features

The stylistic features encode “*how*” humans and bots speak in their interactions. We use two sets of stylistic features: Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) and linguistic accommodation (Danescu-Niculescu-Mizil et al., 2011).

LIWC. We experiment with the psychologically validated word categories (e.g., positive emotion, cognitive, and social processes) in LIWC as features. These lexicons might reveal more about a writer’s thought processes, emotional states, and intentions. For LIWC features, we use the same word categories described in the Linguistic Accommodation Section below to train our LIWC-based classifier. Specifically, we use LIWC to process each utterance. Next, each utterance’s LIWC scores are averaged together to form a dialogue-specific feature vector. We experiment with LIWC features variants: human-only, unknown-only, and human-unknown. Each setting averages the LIWC scores across a different set of utterances, similar to the bag-of-words and embedding features.

Linguistic Accommodation. We use the method

⁴We use the bert-base-uncased pre-trained model available in the HuggingFace package (Wolf et al., 2019).

proposed by Danescu-Niculescu-Mizil et al. (2011) to analyze linguistic alignment. It includes two primary group-level and conversation-level measures: baseline word usage and alignment. The group-level baseline word use corresponds to the rate at which a person uses a given *word category* when it has not been used previously in a conversation. The group-level alignment score reflects the proportionate increase—compared to the baseline score—in the likelihood of the word being used when it has been used previously in the conversation. Similarly, the conversation-level metrics measure accommodation and usage between a human and Unknown (bot or human) user in the each dialogue D_i . We measure the probability of seeing a word category (e.g., LIWC categories) w_c given w_c appeared in the previous utterance $P(w_c \in u_i^o | w_c \in u_{i-1}^h)$ ⁵ and the baseline probability of seeing w_c in the conversation $P(w_c \in u_i^o)$. Next, the accommodation score is calculated by taking the difference between both probabilities, $acc(c) = P(w_c \in u_i^o | w_c \in u_{i-1}^h) - P(w_c \in u_i^o)$. These empirical probabilities are at the conversation level (i.e., measuring how two specific users align). To obtain the group-level estimates, we simply average the accommodation scores $acc(c)$ over all conversations D_i for each word category. We use the following 17 LIWC categories (Pennebaker et al., 2015): i, you, we, they, social, cogproc, posemo, negemo, article, prep, certain, conj, discrep, negate, pronoun, quant, and tentat. We experiment with two main settings as described for the previous methods: Human-Only, Unknown-Only. The Human-Only setting consists of just using the known human’s alignment scores in each conversation and vice-versa for the Unknown-Only setting.

Model Training Details. For the content feature sets, we train a Logistic Regression classifier from the Scikit-Learn package (Pedregosa et al., 2011). Using the validation split for each dataset, we grid-search over the C-values $\{.0001, .001, .01, .1, 1, 10\}$, the logistic regression class weight parameters $\{\text{None, balanced}\}$, and normalization procedures $\{\text{standardize, unit normalize, None}\}$. For the Stylistic features, we train a Random Forest classifier from the Scikit-Learn package. Again, using the validation split, we grid-search

⁵The probabilities are from the perspective of the Unknown user in the conversation. They are also calculated from the human’s perspective.

	U → U	P → P	C → C
Baselines			
Most Frequent	.482	.462	.493
Most Infrequent	.065	.125	.027
Stratified (random)	.491	.498	.487
Content Features			
Human Bag-of-Words	.980	.987	.939
Human BERT	.989	.996	.987
Unknown Bag-of-Words	.971	.970	.997
Unknown BERT	.996	.996	.983
Human + Unknown Bag-of-Words	.958	.963	.986
Human + Unknown BERT	.990	.992	.990
Stylistic Features			
Human LIWC	.878	.899	.705
Human Accommodation	.989	.987	.627
Unknown LIWC	.838	.862	.707
Unknown Accommodation	.887	.897	.674
Human & Unknown LIWC	.885	.903	.601
Human & Unknown Accommodation	.988	.990	.729

Table 2: Source → Source Macro F1 Results for Bot Detection

over the class weight parameters $\{\text{None, balanced, balanced_subsample}\}$, criterion measures $\{\text{gini, entropy}\}$, max features $\{\text{sqrt, log2, None}\}$, and bootstrap parameters $\{\text{True, False}\}$. For all experiments using the Random Forest classifier, we set `n_estimators` to 1000.

5 Results

In this section, we report two sets of results. First, in Subsection 5.1 we present the performance of the bot detection models we explain in Section 4. Beyond the methods described in Section 4, we also compare three baselines: Most Frequent, Most Infrequent, and Stratified. The Most Frequent baseline predicts the most frequent class for every example (i.e., human-human). The Most Infrequent baseline predicts the most infrequent class (i.e., human-bot) and the Stratified baseline makes random prediction proportional to each class’s frequency. Second, in Subsection 5.2, we present a fine-grained analysis of the linguistic accommodation results on the ConvAI2 datasets. For all results, we report the Macro F1 (average F1 for the human-human and human-bot conversation classes).

5.1 Bot Detection Experiments

The source dataset results are reported in Table 2. Overall, we find that content features are the most

		U → C	P → C	C → U	C → P	AVG
<i>Baselines</i>	1. Most Frequent	.493	.493	.482	.462	.483
	2. Most Infrequent	.027	.027	.065	.125	.061
	3. Stratified (Random)	.502	.479	.499	.476	.489
<i>Content: What the bots and humans write?</i>	4. Human bag-of-words	.518	.504	.536	.608	.541
	5. Human BERT	.493	.493	.482	.462	.483
	6. Unknown bag-of-words	.493	.493	.535	.475	.499
	7. Unknown BERT	.493	.493	.482	.462	.483
	8. Unknown and Human bag-of-words	.509	.510	.639	.608	.567
	9. Unknown and Human BERT	.493	.493	.522	.478	.497
<i>Stylistic: How the bots and humans write?</i>	10. Human LIWC	.480	.491	.524	.512	.502
	11. Human Accommodation	.631	.591	.604	.510	.584
	12. Human LIWC + Accommodation	.605	.608	.703	.724	.660
	13. Unknown LIWC	.474	.478	.483	.476	.478
	14. Unknown Accommodation	.428	.424	.503	.494	.462
	15. Unknown LIWC + Accommodation	.462	.436	.502	.497	.474
	13. Human & Unknown LIWC	.521	.504	.532	.556	.528
	14. Human & Unknown Accommodation	.611	.620	.642	.677	.637
	15. Human & Unknown LIWC + Accommodation	.622	.633	.642	.677	.643

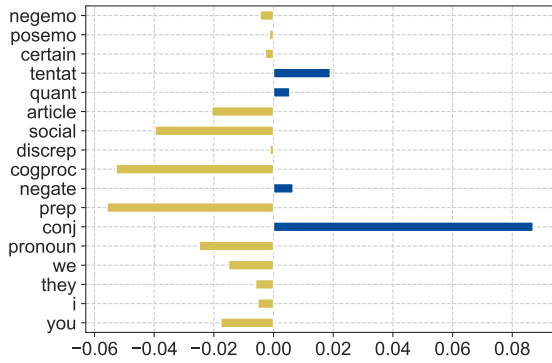
Table 3: This table reports the cross-dataset Macro F1 score for detecting human-bot conversations for three datasets: Unpaid (U), Paid (P), and Control (C). The largest Macro F1 score in each column is in **bold**.

predictive for bot detection, when training and testing on the same train-test splits from the same dataset. Furthermore, we find that the BERT-based models are able to outperform the Bag-of-Words models on average. For instance, the Human Bag-of-Words model on dataset C obtains an F1 of .939. Yet, the Human BERT model obtains an F1 of .987. Likewise, for the stylistic features, we find that simply using LIWC works better than using the accommodation features alone for dataset C. However, combining both Accommodation and LIWC features from both the Human and the Bot is better than using either feature set individually with an F1 of .729.

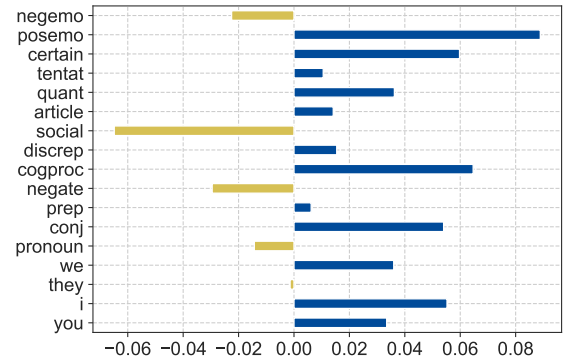
The cross-dataset bot detection results are presented in Table 3. Specifically, the scores are from experiments where we train on a source dataset and evaluate each model on a target dataset’s test split (i.e., source → target). Overall, we make three major findings. First, for both Content and Stylistic features, we find that analyzing the known human’s language in each conversation is more informative than analyzing the bot’s content. For instance, the average (AVG) score for human bag-of-words is .541, while the bot bag-of-words model AVG result is nearly 4% lower (.499). We have similar findings between human LIWC (.502) and bot LIWC (.478). We find that combining both bot and human LIWC improves the AVG performance of the human-only LIWC model with an F1 of 0.528. Second, while the

BERT-based model performs better when applied to data from the same datasets in Table 2 (e.g., U → U) in many settings, when the test dataset changes substantially (e.g., U → C), the generalization performance of BERT drops compared to using bag-of-words. This result is potentially caused by overfitting to random source-specific characteristics. Third, we find that Accommodation features outperform all other individual feature sets with an AVG bot-detection Macro F1 of .584, with the exception of using both the human’s and bot’s accommodation features which has a Macro F1 of .637. Interestingly, human accommodation information (.584) is more predictive than the bot’s text (.462). Intuitively, the type of responses generated by a bot can differ substantially depending on the bot’s training data (e.g., ConvAI2 vs. DailyDialog) and model (e.g., transformer vs. retrieval-based models). Hence, the human’s responses are more consistent with regard to alignment. The best combination overall is the combination of the human’s LIWC features with the human’s accommodation features achieving a Macro F1 of .660. Yet, with the bot (Unknown) feature combination, the performance drops from .768 with Unknown LIWC to .674 after combining accommodation features.

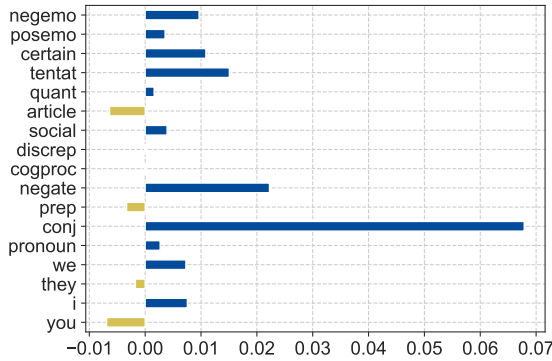
In Table 4, we analyze the Random Forest’s importance scores for the best model (trained on dataset P’s human responses) features using the model that combines human LIWC and accommo-



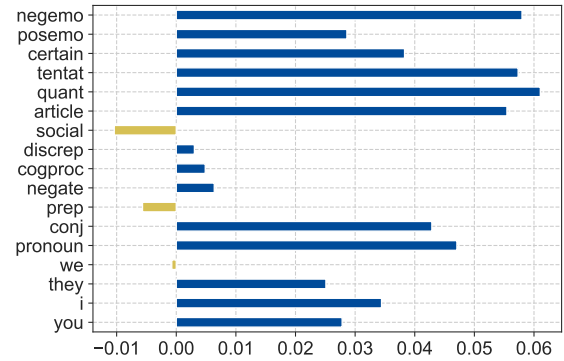
(a) Human TickTok Alignment



(b) Bot TickTok Alignment



(c) Human INTERMEDIATE Alignment



(d) Bot INTERMEDIATE Alignment

Figure 2: Alignment scores for the TickTok and INTERMEDIATE datasets.

Feature Group	Feature	FI
Accommodation	pronoun	.598
LIWC	social	.063
LIWC	article	.043
LIWC	prep	.028
LIWC	pronoun	.020
Accommodation	quant	.018
LIWC	tentat	.017
Accommodation	i	.017
LIWC	you	.016
LIWC	i	.015

Table 4: The top ten most *informative* features for the Random Forest model trained on dataset P’s human responses to detect human-bot conversations with LIWC and Accommodation features.

dation features. The most informative feature is pronoun accommodation. See Section 6 for more details. Other informative features include the human’s use of social, quant (quantitative), and tentative words.

5.2 Accommodation Analysis

In this section, we analyze the group-level (averaged) accommodation scores for the INTER-

MEDIATE and TickTok datasets. Again, note that the group-level scores are dataset averages for each group (human and bot), the conversation-level estimates can vary from the group scores. The accommodation analysis is displayed in Figure 2. The bot results (Figures 2d and 2b) indicate how the bot responds to a human, and the human scores (Figures 2c and 2a) represent how a human responds to a bot. We report all LIWC categories, including Pronoun usage (e.g., pronoun, I, you, we, and they) and Rhetorical words (e.g., social, cogproc, posemo, and negemo) results. Find the analysis for all of the datasets in the Supplementary Material.

Overall, some indicators are robust across datasets, whereas others can vary. For instance, for the general pronoun category, we find that human generally accommodate less towards bots than bots accommodate towards them. The pronoun accommodation pattern for pronouns is robust across both datasets. Similarly, for both datasets, humans accommodate more often than bots for social words. Yet, for negative emotion words (negemo), we find that Bot’s accommodate less than humans

in the TickTock dataset, yet they accommodate more in the INTERMEDIATE dataset. In contrast, in the ConvAI2 and DailyDialog datasets, human accommodation is the same *on average* when two humans participate in a conversation together (not shown in the Figures).

6 Discussion

Why is humans' accommodation such a powerful predictor that they are in a conversation with a bot—even when they are supposedly treating the conversation the same as a human-human interaction? Overall, we found two major reasons to explain the model's high performance. First, modern bots from the transformer to retrieval-based models generate bad responses. For instance, in the conversation below from the VOLUNTEER dataset, the human and bot have already had a conversation about hair color:

```
Bot:   What is your hair color?
Human: I said its black....
Bot:   I know the feeling.
Human: What?
```

From the example, we can see that the bot is asking about hair color, which was already discussed, then the bot responds with a nonsensical output, "I know the feeling." Thus, we see that the final human response does not accommodate with with "I" nor any first singular pronoun. The lower alignment score for humans (compared to bots) for the word category I in Figure 2 supports this finding. Furthermore, from Table 4, we know that accommodation in terms of pronoun usage is highly predictive of a human-bot conversation. Similar posts can be found in all datasets where bad outputs can explain why humans align differently concerning "I" in bot-human interactions than human-human interactions. This finding is interesting because it is similar to the findings in human-human interactions where alignment is positively correlated with understanding in cooperative decision-making tasks (Fusaroli et al., 2012; Kacewicz et al., 2014). If humans don't understand the conversation, they align less, at least with regards to pronoun usage in our case. The next example from the VOLUNTEER dataset is a better, more coherent conversation:

```
Human: What is your Hobby?
Bot:   I love music! I love Taylor Swift
Human: Me too. And what about Iggy Pop?
```

We can see that the final human response aligns with the first person singular word "me" to the

bot's use of "I". A potential use that warrants future work in the area of bots evaluation involves using accommodation information as an additional evaluation metric in human evaluation studies if certain types of accommodation correlate with bad system responses. If human ratings of bot performance are noisy, the use of accommodation information can provide insight into how well a bot performs, if a human interacts with it differently than they do in human-human conversations.

The second reason for the significant predictive performance can be seen in adversarial conversations between the human and bot. Specifically, in instances where the human is trying to trick the bot or simply trying to ask questions, bot-human alignment can differ from human-human conversations. The example below from the TickTock dataset

```
Human: Are you a robot?
Bot:   No.
Human: Where are you from?
```

shows repeated "You" usage by the user, repeatedly asking multiple questions. Again, differences in pronoun usage can reflect adversarial users. Again, in terms of bot evaluation, if a human is acting adversarially, and it is possible to differentiate adversarial interactions from the human not understanding, this can provide insight into . More examples can be found in the Supplementary Material.

7 Conclusion

This study has is an exploration of the levels of linguistic alignment in human-bot conversations and presents evidence about how bot-human interactions are different than human-human conversations. This work's insights have implications for future bot detection methodologies. How humans and bots come together and converse manifested by their level of agreement and disagreement is a first step towards understanding human interactions with general conversational agents. Furthermore, as future work, we will explore the use of linguistic alignment to improve bot evaluation frameworks involving human-bot interactions. Moreover, we will also explore methods of controlling the accommodation level of bots themselves. Given the implicit social implications of certain types of accommodation, controlling it can help important tasks (e.g., counseling or customer service-related bots).

References

- Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan Black. 2020. What code-switching strategies are effective in dialogue systems? *Proceedings of the Society for Computation in Linguistics*, 3(1):308–318.
- Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42.
- David M Beskow and Kathleen M Carley. 2018. Bot conversations are different: leveraging network metrics for bot detection in twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 825–832. IEEE.
- Ana Paula Chaves. 2020. Should my chatbot be register-specific? designing appropriate utterances for tourism. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11.
- Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It’s how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 102–109.
- Michelle Cohn, Chun-Yen Chen, and Zhou Yu. 2019. A large-scale user study of an alexa prize chatbot: Effect of its dynamism on perceived quality of social dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 293–306.
- Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754.
- Florian Daniel, Cinzia Cappiello, and Boualem Benaallah. 2019. Bots acting like humans: Understanding and preventing harm. *IEEE Internet Computing*, 23(2):40–49.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (conva2). *arXiv preprint arXiv:1902.00098*.
- Gabriel Doyle and Michael C Frank. 2016. Investigating the sources of linguistic alignment in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 526–536.
- Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 301–305.
- Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. 2012. Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological science*, 23(8):931–939.
- Andres Garcia-Silva, Cristian Berrio, and José Manuel Gómez-Pérez. 2019. An empirical study on pre-trained embeddings and language models for bot detection. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 148–155.
- Howard Giles, Donald M Taylor, and Richard Bourhis. 1973. Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in society*, pages 177–192.
- Augusto Gnisci. 2005. Sequential strategies of accommodation: A new method in courtroom. *British Journal of Social Psychology*, 44(4):621–643.
- Jonathan Herzig, Tommy Sandbank, Michal Shmueli-Scheuer, and David Konopnicki. 2019. Bot2vec: Learning representations of chatbots. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 75–84.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.
- Ganesh Jawahar, Muhammad Abdul-Mageed, and VS Laks Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309.
- Simon Jones, Rachel Cotterill, Nigel Dewdney, Kate Muir, and Adam Joinson. 2014. Finding zelig in text: A measure for normalising linguistic accommodation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 455–465.
- Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.

- Dominika Kaczorowska-Spychalska. 2019. How chatbots influence marketing. *Management*, 23(1):251–270.
- Alice Kerlyl, Phil Hall, and Susan Bull. 2006. Bringing chatbots into education: Towards natural language negotiation of open learner models. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 179–192. Springer.
- Susan L Kline and Janet M Ceropski. 1984. Person-centered communication in medical practice. *Human decision-making*, pages 120–141.
- Jürgen Knauth. 2019. Language-agnostic twitter-bot detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 550–558.
- Naomi Kong-Vega, Mingxin Shen, Mo Wang, and Luis Fernando D’Haro. 2019. Subjective annotation and evaluation of three different chatbots wochat: Shared task report. In *9th International Workshop on Spoken Dialogue System Technology*, pages 371–378. Springer.
- Dijana Kosmajac and Vlado Keselj. 2019. Twitter bot detection using diversity measures. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 1–8.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.
- Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2020. Learning healthbots from training data that was automatically created using paraphrase detection and expert knowledge. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 638–648.
- Ming-Cheng Ma and John P Lalor. 2020. An empirical analysis of human-bot interaction on reddit. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 101–106.
- Alistair Martin, Jama Nateqi, Stefanie Gruarin, Nicolas Munsch, Isselmou Abdrahmane, Marc Zobel, and Bernhard Knapp. 2020. An artificial intelligence-based first-line defence against covid-19: digitally screening citizens for risks via a chatbot. *Scientific reports*, 10(1):1–7.
- Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.
- Timothy Niven and Hung-Yu Kao. 2019. Detecting argumentative discourse acts with linguistic alignment. In *Proceedings of the 6th Workshop on Argument Mining*, pages 104–112.
- David Oniani and Yanshan Wang. 2020. A qualitative evaluation of language models on automatic question-answering for covid-19. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9.
- Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. 2020. Detection of bots in social media: a systematic review. *Information Processing & Management*, 57(4):102250.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.
- Roberto Pieraccini, David Suendermann, Krishna Dayanidhi, and Jackson Liscombe. 2009. Are we there yet? research in commercial spoken dialog systems. In *International Conference on Text, Speech and Dialogue*, pages 3–13. Springer.
- Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.
- Hagyeong Shin and Gabriel Doyle. 2018. Alignment, acceptance, and rejection of group identities in online political discourse. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 1–8.
- Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.
- Paul J Taylor and Sally Thomas. 2008. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, 1(3):263–281.
- Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1171–1180.
- Yafei Wang, David Reitter, and John Yen. 2014. Linguistic adaptation in conversation threads: Analyzing alignment in online health communities. *ACL 2014*, page 55.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Nathan I Wood. 2019. Departing from doctor-speak: A perspective on code-switching in the medical setting. *Journal of general internal medicine*, 34(3):464–466.

Yang Xu, Jeremy Cole, and David Reitter. 2018. Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 601–610.

Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2019. Feedpal: Understanding opportunities for chatbots in breastfeeding education of women in india. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30.

Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. 2015. Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness. In *2015 AAAI Spring symposium series*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

A Expanded Alignment Analysis

The complete group-level alignment scores for the IRIS and Tolokers datasets can be found in Figures 3a, 3b, 3c, and 3d.

B Example Conversations

Here we present examples of conversations between humans and bots that are representative of good conversational structure in terms of alignment between the two parties such as the usage of pronoun “I” in the first example.

Bot: I love iphone!
i just bought new iphone!
Human: Thats good for you i m not
very into new tech

Human: I am fine. I like hip hop.
What about you?
Bot: I like r b and pop
are you 100 years old?
Human: What? No. You are rude

Human: Do you have dogs?
Bot: Yes i have two dogs and a cat.
Human: Cool I love dogs.
But i have siblings instead...
Bot: I have two brothers and
a sister and a sister.

Here conversations between humans and bots involve irregular pronoun usage that are representative of a break in the conversational pattern and little or no alignment between the two parties such as the usage of pronouns “You” and resultant use of “I” reply in the first example.

Human: See you then!
Bot: I m not going to see you.

Bot: I am 24 and I have a dog named
her name is named name
Human: Really her name is name?
Bot: She is a good name.

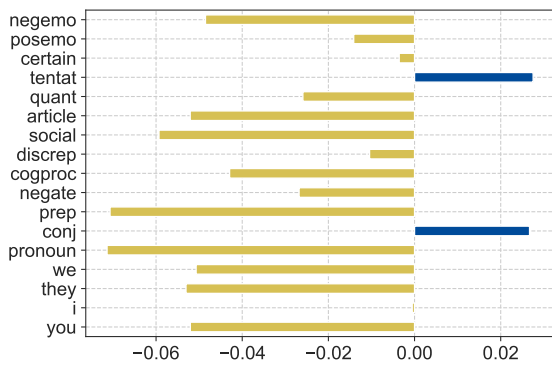
Human: who is obama
Bot: i partially have interest in that
Human: can you tell me more about obama
Bot: i hate pigs i live a farm

These conversations signify a complete breakdown of conversational structure with respect to alignment, as there is none. The chats seem haphazard and vague with either the bot or the human replying in an ambiguous manner.

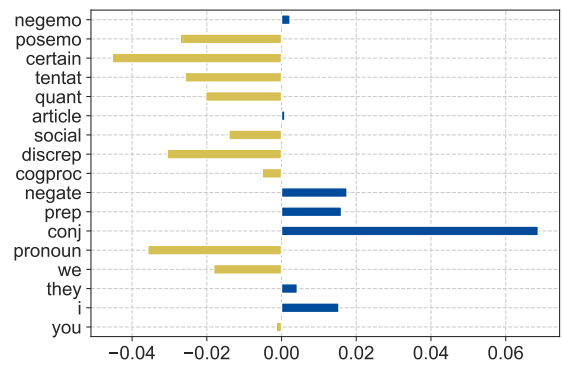
User: Good afternoon
System: Well you know I’m not sure.
What do you think about movies
User: That didn’t make any sense

User: hi again
System: Of course. Any thoughts?
User: no

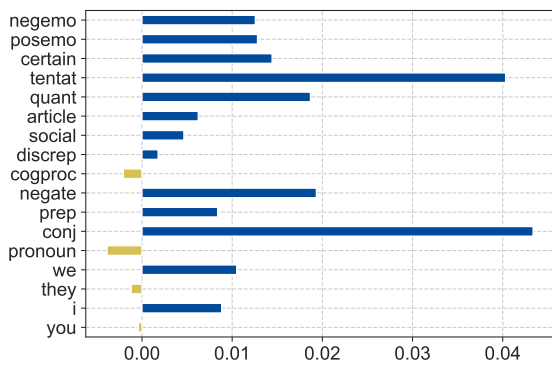
User: Hello
System: Do not argue with an idiot.
He will drag you down to his
level and beat you with
experience.
User: Um Okay that was random



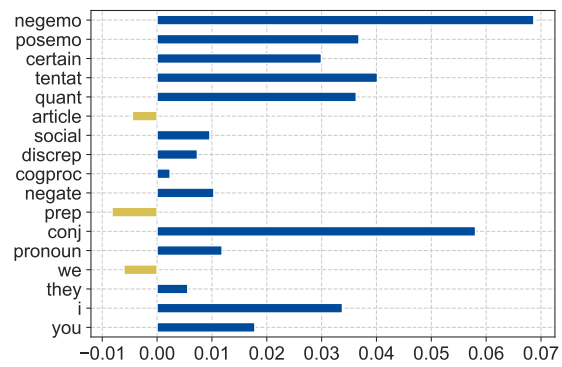
(a) Human IRIS Alignment



(b) Bot IRIS Alignment



(c) Human Tolokers Alignment



(d) Bot Tolokers Alignment

Figure 3: Alignment scores for the IRIS and Tolokers dataset