

Quantifying trust towards LLM-based chatbots: A mixed-method approach

Belosevic, Milena

milena.belosevic@uni-bielefeld.de
Bielefeld University, German Linguistics/Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University, Germany

Buschmeier, Hendrik

hbuschme@uni-bielefeld.de
Bielefeld University, German Linguistics/Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University, Germany

1 Introduction

Trust plays a crucial role in human interaction and has therefore been approached from sociological, psychological, philosophical, and the perspective of communication studies (Hendriks et al., 2021). Given the current popularity of large language model/LLM-based chatbots that can interact in a human-like, conversational way (Rudolph and Samson, 2023), it can be stated that trust in automation is gaining increasing importance. Similar to social norms such as politeness (Lumer and Buschmeier, 2022) in human-machine interaction, trust in automation (Lee and See, 2004, Lukyanenko et al., 2022 for an overview) can also be regarded as a shared social value constructed in direct or indirect interaction. However, in linguistic research on trust (Schäfer, 2016, Belosevic 2022), this aspect has rarely been addressed (Schneider et al., 2022, Lotze, 2016, Kabir et al. 2023).

Since trust comprises both cognitive and emotional aspects (Kok and Soh, 2020), this paper focuses on the role of emotional aspects of trust in indirect interaction with chatbots and uses the perceived trustworthiness ascribed to ChatGPT (as one of the most recent LLM-based chatbots) as a testbed. Since little research has focused on how language shapes the evaluation of trustworthiness ascribed to machines in their role as trust objects, the paper aims to show how trust in human interaction with chatbots can be modeled using manual annotation and sentiment analysis. This stands in contrast to recent studies on the role of trust in ChatGPT, which are mainly based on experimentally elicited data and do not consider the role of language (e.g., Funke et al., 2023, Shen et al. 2023, Watters and Leman-ski, 2023, Huang et al. 2023, Liu et al. 2023). In particular, we propose a mixed-method approach to quantify the trustworthiness ascribed to ChatGPT in an indirect interac-

tion. In this interaction mode, the distinction between the first- and third-person perspective in the human-machine interaction (Coeckelbergh, 2011) is crucial. Whereas the first-person perspective is concerned with how we interact with chatbots, the third-person perspective is adopted in this paper. It explores how users talk about chatbots and how the perceived trustworthiness of ChatGPT is promoted through the discursive commodification of trust (cf. Krüger and Wilson, 2022) in the public debate about ChatGPT in Germany. To this end, we use qualitative approaches to trust (identification of trust-relevant vocabulary through manual annotation) and qualitative-quantitative methods (sentiment analysis) to account for the emotional aspects of trust in human-chatbot interaction.

Prior to applying these methods to our case study, it is necessary to define trust in automation and specify the properties of trust underlying the annotation scheme and their relation to sentiment values.

2 Methodology and data

Trust is a complex phenomenon that can be operationalized using other more concrete concepts (so-called trust cues or trust indicators) as a proxy. This is also true for trustworthiness (cf. Lewicki and Alister, 1998). To identify emotional aspects of trustworthiness, linguistic units that serve as indicators of perceived trustworthiness must first be detected. We consider manual annotation to be the first step toward modeling linguistic indicators of trustworthiness and narrowing down the complex concept of trust into more concrete aspects. The annotation task is based on an annotation scheme with several annotation categories defined by drawing on existing studies on trust in human-human interaction (cf. Kuhnhehn 2014) and human-robot interaction (cf. de Visser et al. 2020)¹. The central annotation unit, namely the notion of perceived trustworthiness was adopted from the concept of trust calibration (Lee and See, 2004, Muir, 1994) which is an often applied framework in studies on trust in human-machine interaction (cf. Wischnewski et al. 2023 for an overview). For the purposes of the annotation scheme, it was defined as the perception of users' trust toward the system in contrast to the actual trustworthiness of the trust object.

Based on the results of the manual annotation, sentiment analysis was carried out to account for the role of emotional aspects of perceived trustworthiness in the public discussion about ChatGPT. We tested one machine-learning-based and one lexicon-based model for sentiment analysis of German texts: the model for sentiment classification available on Hugging Face 3 (pre-trained on 1.834 million German-language samples, mainly texts from Twitter, Facebook, movie, app, and hotel reviews, Guhr et al., 2020) and the Python package textblob (Loria, 2020) based on the German polarity lexicon. The German version of textblob can be used to obtain polarity ratings between -1 (negative) and +1 (positive) for words, sentences, and texts. Both the machine-learning-based model and textblob provide ratings

based on sentences, phrases, and single words. Trust-related linguistic markers were also annotated manually with regard to the promotion of trust or distrust.

The data (27.138 tokens) were obtained from the DWDS-Webmonitor corpus² using the word *ChatGPT* and the period between 2022-11-30 (release of ChatGPT) and 2023-06-30. This dataset comprises some 6.198.349 texts (mostly web pages from German-speaking countries). However, only the intermediate sentence context comprising the search word is available for analysis.

3 Manual annotation and sentiment analysis

As mentioned above, the central part of manual annotation includes the development of an annotation scheme and the definition of annotation categories based on the definition of the main aspects of trust provided in the previous section. The annotation scheme consists of the following annotation categories: interaction mode, trust levels and sentiment values, trust roles in human interaction with ChatGPT, and perceived trustworthiness.

The data were annotated by one annotator using the software MAXQDA Plus³. Two types of annotation categories were combined: the annotation with indicators of trust on the word, multi-word, and sentence level as well as sentiment values. Since indicators of trust comprise several aspects (ability, benevolence, and integrity) the annotation of these aspects in terms of positive and negative sentiment values (trustworthiness vs. distrust) is closely related to aspect-based sentiment analysis that goes beyond the formal level of single words and sentences and focuses on properties of aspect categories (cf. Liu 2015: ch. 5 and 6). In the following, the annotation scheme will be described.

For the annotation category ‘interaction mode’, we annotated the following domains in which the perceived trustworthiness ascribed to ChatGPT is discussed: education, science, politics, sports, and industry. To determine the mode, the annotator often checked the whole text in which the example appeared.

Trust roles comprise the society and users in their roles as trustors on the one side and trust objects (here ChatGPT) on the other. Since our case study is concerned with the public debate about perceived trustworthiness ascribed to ChatGPT by the users (i.e., during the interaction), the central aspect of this annotation unit is not the user, but the discourse actors (e.g., journalists, experts) who indicate their perception of the perceived trustworthiness of users.

To identify linguistic cues of perceived trustworthiness, we draw on the linguistic markers of credibility and trustworthiness proposed by Kuhnhehn (2014) and Reinmuth (2006) for human-human interaction, the categories of trust in human-robot interaction (cf. de Visser et al. 2020), and on the three properties of trust (Mayer et al., 1995) widely accepted in the literature, namely, competence (defined as skills, and characteristics that enable the trustee to influence

the domain), benevolence (specified as the extent to which the intents and motivations of the trustee are aligned with those of the trustor), and integrity (the degree to which the trustee adheres to a set of principles the trustor finds acceptable). Each linguistic marker was annotated with one of the indicators of trustworthiness (competence, benevolence, or integrity). Manual annotation is necessary as there is no agreement about which linguistic units can be regarded as trust-relevant. Moreover, for each domain in which the role of trust is investigated, trust-relevant aspects should be defined based on indicators underlying the construction of trust in the context.

The annotation category ‘trust levels/sentiment values’ is based on our hypothesis that emotional aspects of trust are related to positive emotions and vice versa so that sentiment values can be regarded as a potential cue of emotional dimensions of trustworthiness. Therefore, the levels of positive trustworthiness, negative trustworthiness (distrust), and ambiguous cases were considered sentiment values and served as annotation units for this category. Specifically, trustworthiness was annotated with 1, negative trustworthiness/distrust with -1, and in cases where there was no clear distinction regarding the sentiment value ‘both/ambiguous’ was annotated with 0. To ensure that only trust-relevant aspects and not general emotional aspects are considered for the sentiment analysis, only aspects previously annotated with linguistic trust cues of perceived trustworthiness were annotated with sentiment values. However, we remain agnostic about the exact relation between sentiment analysis and trust(worthiness) because trust comprises further aspects, such as cognitive and attitudinal properties that cannot be completely captured through sentiment analysis and require consideration of further methods.

To illustrate how the annotation scheme was implemented in our dataset, consider the following example extracted from the DWDS WebXL subcorpus:

1. Auch die eloquenten, teilweise charmanten Antworten, die ChatGPT auf bestimmte Fragen gibt, sind manchmal nicht mehr als plausibel klingende Unwahrheiten – man spricht dann davon, so Horn, dass die KI "halluziniert".

‘Even the eloquent, often witty answers that ChatGPT provides to some questions are sometimes nothing more than plausible-sounding untruths – according to Horn, the AI is said to hallucinate.’

In each example, the aspects of trustworthiness (competence, benevolence, and integrity) were annotated with linguistic markers of each category according to the categorization provided in previous studies (Kuhnhehn 2014, Reinmuth 2006, de Visser et al. 2020). Afterward, the linguistic markers were annotated with trust levels/sentiment values. In example (1) the adjectives *eloquent*, *charmant*, nominal phrase *manchmal nicht mehr als plausibel klingende Unwahrheiten* and the verb *halluziniert* were identified as trust-relevant vocabulary. Next, they were annotated with positive trustworthiness (*eloquent*, *charmant*), and

distrust/negative trustworthiness (*manchmal nicht mehr als plausibel klingende Unwahrheiten* und *halluziniert*). In a further step, the example was considered as distrust. In addition to linguistic cues of trustworthiness and sentiment values, trust roles and the interaction mode were annotated separately. In this case, the interaction mode is 'industry' (based on the information provided in the full text⁴), trust roles include ChatGPT as a trust object, and [Dennis] Horn as a trustor. Further examples can be found in the annotation guidelines.

In the next step, we focus on the correlation between trust-related vocabulary obtained by manual annotation and its sentiment values obtained by human sentiment ratings, lexicon-based, and machine-learning-based sentiment models. Positive sentiment scores are related to trustworthiness and vice versa: negative sentiment scores should be related to the erosion of trustworthiness. Neutral scores indicate that both a decrease and increase in trustworthiness can be observed or that there are no sentiment scores. Human sentiment ratings are based on the manual annotation described above. The words, multi-word units, and sentences annotated with human sentiment ratings were imported into Python to obtain their sentiment scores using machine-learning-based and lexicon-based models. We compared the distribution of human ratings with the sentiment scores provided by the pre-trained model (cf. Guhr et al., 2020) and the sentiment analyzer provided in the German language extension for textblob⁵.

4 Results

The annotation with the categories *trust*, *distrust*, and *both/ambiguous* yields that the promotion of trustworthiness occurs more frequently (57.24 %) than the lack of trustworthiness towards ChatGPT. Questions in which trust-relevant aspects could not be identified in the context (e.g., "Ist ChatGPT kostenlos?") were excluded from the analysis. The analysis indicates that ca. 75 % of data accounts for the aspect *competence*, usually regarding how ChatGPT can be trusted to provide users with accurate information and ensure that the provided information is reliable. The manual annotation yielded some 6480 trust-relevant linguistic markers on the word-, multiword-, and sentence level (25 % of the total number of tokens) that were selected for further analysis. They comprise trust-relevant vocabulary annotated within each trust-relevant utterance.

Regarding the results of the sentiment analysis, the annotated words, multi-word units, and sentences were imported into Python to obtain their sentiment scores. The scores obtained by the trained model are negative (40 %) or neutral (38 %), and only 20 % of the annotated data are positive. Human ratings are 45 % negative and 52 % positive, less than 1 % was rated as neutral. As compared to human and machine-learning-based ratings, the majority of ratings obtained by textblob are neutral. In particular, textblob rated 11 % of trust-related vocabulary as negative and

25.8 % as positive. The results indicate significant differences in sentiment scores between human-, lexicon-based, and machine-learning-based ratings, especially regarding the amount of neutral ratings in non-human-based models. Regarding the correlation between sentiment scores and human-based evaluation of trustworthiness, preliminary results indicate a higher correlation between negative sentiments and the lack of trustworthiness and vice versa between positive sentiment scores and the assignment of trustworthiness.

5 Conclusions and outlook

The paper explored indirect measures of emotional aspects of trust that go beyond linguistic units such as trust, mistrust, or trustworthiness and, in contrast to direct measures (e.g., scales), require a qualitative approach as a first step toward detecting the role of trust in a particular context. Since indirect measures are highly dependent on human interpretation, they pose a challenge to the research on trust and put the objectivity of qualitative measures into question. In this paper, we argued that Digital Humanities offers appropriate methods to remedy these issues.

The results show how qualitative and quantitative methods used in the Digital Humanities contribute to studies on trust in human-machine interaction. On the other hand, trust in automation as the object of investigation contributes to ongoing debates regarding the reliability of sentiment measures for languages other than English (Kaity and Balakrishnan, 2020) and provides empirical evidence for how sentiment scores can be used for modeling social phenomena like trust.

Fußnoten

1. The annotation guidelines are available online at <https://doi.org/10.17605/OSF.IO/FVB7P>.
2. <https://www.dwds.de/d/korpora/webxl>
3. <https://www.maxqda.com/>
4. <https://web.archive.org/web/20230616185754/https://www.boersenblatt.net/news/boersenverein/digitaler-wandel-nachhaltig-gedacht-289685>
5. <https://textblob-de.readthedocs.io/>

Bibliographie

- Belosevic, Milena.** 2022. *Vertrauen und Misstrauen in der Flüchtlingsdebatte 2015-2017. Eine diskurslinguistische Untersuchung von Argumentationsmustern*. Hamburg: Buske. <https://doi.org/10.46771/978-3-96769-198-6>.
- Coeckelbergh, Mark.** 2011. „You, Robot: On the Linguistic Construction of Artificial Others“. *AI & SOCIETY* 26 (1): 61–69. <https://doi.org/10.1007/s00146-010-0289-z>.

- De Visser, Ewart J., Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx.** 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12 (2): 459–78. <https://doi.org/10.1007/s12369-019-00596-x>.
- Farhat, Faiza, Shahab Saquib Sohail, und Dag Øivind Madsen.** 2023. „How Trustworthy is ChatGPT? The Case of Bibliometric Analyses“. Preprint. Social Sciences. <https://doi.org/10.20944/preprints202303.0479.v1>.
- Funke, Noemi, Katja Stadler, Heidi Vakkuri, Anna Wagner, Marc Lunkenheimer, und Alexander H. Kracklauer.** 2023. „Your Conversational Partner Is a Chatbot“ - An Experimental Study on the Influence of Chatbot Disclosure and Service Outcome on Trust and Customer Retention in the Fashion Industry.“ <https://doi.org/10.25929/JAIR.VIII.113>.
- Guhr, Oliver, Anne-Kathrin Schumann, Frank Bahrmann, und Hans Joachim Böhme.** 2020. „Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems“. In *LREC 2020 Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Palais Du Pharo, Marseille, France: Conference Proceedings*, herausgegeben von Nicoletta Calzolari, 1627–32. Paris: ELRA - European Language Resources Association. <https://aclanthology.org/2020.lrec-1.202>.
- Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. De Visser, und Raja Parasuraman.** 2011. „A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction“. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53 (5): 517–27. <https://doi.org/10.1177/0018720811417254>.
- Hendriks, Friederike, Bettina Distel, Katherine M. Engelke, Daniel Westmattmann, und Florian Wintterlin.** 2021. „Methodological and Practical Challenges of Interdisciplinary Trust Research“. In *Trust and Communication*, herausgegeben von Bernd Blöbaum, 29–57. Cham: Springer. https://doi.org/10.1007/978-3-030-72945-5_2.
- Huang, Xiaowei, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, et al.** 2023. “A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation.”
- Kabir, Samia, David N. Udo-Imeh, Bonan Kou, und Tianyi Zhang.** 2023. “Who Answers It Better? An in-Depth Analysis of ChatGPT and Stack Overflow Answers to Software Engineering Questions”. arXiv. <https://doi.org/10.48550/arXiv.2308.02312>.
- Kaity, Mohammed, und Vimala Balakrishnan.** 2020. „Sentiment Lexicons and Non-English Languages: A Survey“. *Knowledge and Information Systems* 62 (12): 4445–80. <https://doi.org/10.1007/s10115-020-01497-6>.
- Kok, Bing Cai, und Harold Soh.** 2020. „Trust in Robots: Challenges and Opportunities“. *Current Robotics Reports* 1 (4): 297–309. <https://doi.org/10.1007/s43154-020-00029-y>.
- Krüger, Steffen, und Christopher Wilson.** 2023. „The Problem with Trust: On the Discursive Commodification of Trust in AI“. *AI & SOCIETY* 38 (4): 1753–61. <https://doi.org/10.1007/s00146-022-01401-6>.
- Kuhnhehn, Martha.** 2014. *Glaubwürdigkeit in der politischen Kommunikation Gesprächsstile und ihre Rezeption*. Konstanz; München: UVK-Verl.-Ges.
- Lee, J. D., und K. A. See.** 2004. „Trust in Automation: Designing for Appropriate Reliance“. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.
- Lewicki, Roy J., Daniel J. McAllister, and Robert J. Bies.** 1998. „Trust and Distrust: New Relationships and Realities“. *The Academy of Management Review* 23 (3): 438. <https://doi.org/10.2307/259288>.
- Liu, Yang, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li.** 2023. “Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models’ Alignment.” <https://doi.org/10.48550/ARXIV.2308.05374>.
- Liu, Bing.** 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge Univ. Press.
- Loria, Steven.** 2020. „textblob Documentation. Release 0.16.0“. <https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf>.
- Lotze, Netaya.** 2016. *Chatbots*. Bern: Peter Lang. <https://doi.org/10.3726/b10402>.
- Lukyanenko, Roman, Wolfgang Maass, und Veda C. Storey.** 2022. „Trust in Artificial Intelligence: From a Foundational Trust Framework to Emerging Research Opportunities“. *Electronic Markets* 32 (4): 1993–2020. <https://doi.org/10.1007/s12525-022-00605-4>.
- Lumer, Eleonore, und Hendrik Buschmeier.** 2022. „Modeling Social Influences on Indirectness in a Rational Speech Act Approach to Politeness“. In *Proceedings of the 44th Annual Conference of the Cognitive Science*, herausgegeben von Jennifer Culbertson, Andrew Perfors, Hugh Rabagliati, und Veronica Ramenzoni, 2796–2802. Toronto.
- Mayer, Roger C., James H. Davis, and F. David Schoorman.** 1995. „An Integrative Model of Organizational Trust“. *The Academy of Management Review* 20 (3): 709. <https://doi.org/10.2307/258792>.
- Muir, Bonnie M.** 1994. „Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems“. *Ergonomics* 37 (11): 1905–22. <https://doi.org/10.1080/00140139408964957>.
- Reinmuth, Marcus.** 2006. *Vertrauen schaffen durch glaubwürdige Unternehmenskommunikation - Von Geschäftsberichten und den Möglichkeiten und Grenzen einer angemessenen Sprache*. Dissertation. Düsseldorf. <https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=3547>

Rudolph, Jürgen, and Tan Samson. 2023. „ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?“ *Journal of Applied Learning & Teaching* 6 (1). <https://doi.org/10.37074/jalt.2023.6.1.9> .

Schäfer, Pavla. 2016. *Linguistische Vertrauensforschung: Eine Einführung* . Berlin: De Gruyter. <https://doi.org/10.1515/9783110451863> .

Schneider, Britta, Bettina Migge, Doris Dippold, Iker Erdocia, Marie-Theres Fester-Seeger, Sviatlana Höhn, Ledia Kazazi, u. a. 2022. „Changing Language Ideological Concepts in the Human-Machine Era. Questions, Themes and Topics“. <https://doi.org/10.13140/RG.2.2.25867.36649> .

Shen, Xinyue, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. „In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT“. <https://doi.org/10.48550/ARXIV.2304.08979> .

Watters, Casey and Michal Lemanski. 2023. „Universal Skepticism of ChatGPT: A Review of Early Literature on Chat Generative Pre-Trained Transformer“. *Frontiers in Big Data* , Nr. 6.

Wischnewski, Magdalena, Nicole Krämer, und Emmanuel Müller. 2023. „Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-of-the-Art and Future Directions“. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* , 1–16. Hamburg Germany: ACM. <https://doi.org/10.1145/3544548.3581197> .