To what extent do uncertainty, sentiment, and authors influence the amount of anthropomorphic behavior on social media?

THOM HARBERS, University of Twente, The Netherlands

Social media serves as a platform for people to share their thoughts but also provides a place to share knowledge. For certain accounts, people tend to attribute human-like qualities to non-human agents. We are interested in seeing what factors and strategies influence the amount of anthropomorphic behavior elicited, with the goal of creating social entities. We performed an analysis of three different Twitter accounts, which we labeled with different message composition strategies. By analyzing the replies to a multitude of tweets for each account, we examined whether uncertainty and sentiment in a non-human agent's behavior heightened the amount of anthropomorphic behavior elicited by consumers. A large dataset was investigated using natural language processing techniques. Anthropomorphism was measured by examining the use of pronouns within replies, as these contained a lot of first-person ("I"), second-person ("you"), and third-person pronouns ("he/she"), as well as explicit mentioning of authors' names. Uncertainty and sentiment of a non-human agent's behavior resulted in insignificant results with no correlation between them and anthropomorphism. Our findings do suggest that there is a significant difference between the authors from the accounts analyzed (human written vs. computer generated), however, there may have been additional or different factors to the ones analyzed that resulted in this.

Additional Key Words and Phrases: social media, anthropomorphism, animacy, natural language processing, linguistic analysis

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Bob Schadenberg, and my temporary supervisor Dr. Lorenzo Gatti, for their guidance throughout the research project. Their insights were vital in shaping the direction of this project.

TScIT 39, July 7, 2023, Enschede, The Netherlands

1 INTRODUCTION

The use of sensors and other technologies has allowed farmers to collect and analyze a wide variety of data regarding their farms. With this data, farmers can gain insights into the health and productivity of plants and animals, as well as the environmental impact of the farming operation. Using this information, areas that need attention can quickly be identified and efficiently be helped, as the impact can easily be monitored. These advanced technologies have attracted the attention of an increasing number of farmers. They are interested in getting to know as much as they can about their farm, ultimately resulting in better decision-making and results. While they are increasingly gaining information, the general public, who are not involved in the agricultural sector, remains uninformed about modern farming practices. A reason for this is the inaccessibility of information, resulting in a gap between producer and consumer.

Researchers have looked into bridging this gap. Duffy et al. [5] came to the conclusion that unless organizations have significant funding to get their message across to consumers, people will not be interested in thinking about the environment, animal welfare, or local consumers. Especially individual organizations that do wish to communicate with their target audience face difficulties, as they do not have the financial capabilities. The pooling of resources and collaboration on a communication strategy seems possible, however, is difficult, as organizations have conflicting approaches to food production. Since the publication of this study, the media landscape has undergone significant changes. The rise of social media potentially offers a cost-effective and powerful platform for organizations, compared to traditional media. Additionally, with sensor technologies being utilized more often, a combination of these two could play a role in the process of trying to bridge the gap. The process of creating social entities by eliciting anthropomorphic behavior could be a promising way to communicate. Anthropomorphism is described as attributing human-like mental states, feelings, and characteristics to objects, animals, and other phenomena [10]. This approach can be used in hopes to make the information more comprehensible [10], as social entities possibly enable farmers to create a more personal connection with consumers. This could allow farmers to explain their practices by telling an informative and compelling story, while also doing it in an interesting and easily accessible way.

The problem is that it is not clear how a social entity (e.g. farm, animal, or robot) should communicate. Therefore, our goal is to see what effect certain factors and authors have on anthropomorphic communication on social media. To measure these effects, we will be looking at the pairings of tweets from several authors, and the use of pronouns by consumers in the replies. Epley et al. [6] suggest that three factors make up the psychological phenomenon of anthropomorphism, with one of these factors relating to uncertainty and unpredictability. Therefore, the correlation between the similarity of an agent's messages and the language used by consumers might provide useful insights. Additionally, it has been shown that emotional feedback of an agent influences the intent to interact with it [14], making the sentiment of messages and the relationship to consumer's replies interesting to look at. There are also different ways a message from a social entity could be composed. To see the effect, we will be looking at the author of the tweets, specifically by comparing tweets written by humans imposing as a social entity (ghostwritten), and computer-generated tweets.

To achieve our goal we will use research questions (RQ) as a basis of our research.

- RQ 1: To what extent does similarity between a non-human agent's messages influence the number of pronouns used by consumers?
- RQ 2: How does the sentiment of a non-human agent's messages correlate with the number of pronouns used by consumers?
- RQ 3: To what extent does the author affect the use of pronouns in consumer replies?

By the end of this research, we hope to have contributed in two different ways. Firstly, we aim to extract factors that are effective in current anthropomorphic communication by analyzing the interaction of the general public with such messages. Secondly, we hope to have discovered what the effect is of an author on the amount of anthropomorphic behavior displayed.

In this document, we will first look at related works regarding anthropomorphic communication in section 2. In section 3 will describe the methodologies used to answer the research questions. Within section 4 we will report on the results achieved after performing the described methods. These results will be discussed in section 5, in which limitations of the methods used will also be covered, as well as ideas for opportunities to further expand upon this work. Lastly, section 6 will show the conclusions from our research.

2 RELATED WORKS

Researchers have been looking into anthropomorphism for the past couple of decades now. Caporael [1] has suggested that anthropomorphism should be regarded as a psychological phenomenon on its own. Especially as a "default schema under conditions of quasi-predictability", where computers and robots are viewed similarly to humans [2].

One of the larger studies in the domain of anthropomorphism is one by Epley et al. [6]. They also describe anthropomorphism as a psychological phenomenon, which they constructed three factors around. Firstly, *elicited agent knowledge* describes how knowledge about human and non-human agents is activated, acquired, corrected, and applied to a target. Reasoning about other humans using one's own mental states and characteristics as a guide is egocentrism. When reasoning about a non-human agent, this same process would describe anthropomorphism. However, the degree to which this applies varies. It is intuitive for people to use one's self-knowledge as a starting point for reasoning about non-human agents, but it is possible to correct this view if they are motivated to do so. The extent to which people are willing to expend additional resources can be described as the need for cognition. People who are high in need often enjoy engaging in effortful thinking and are more likely to overcome default assumptions, therefore showing weaker elicitation of anthropomorphic behavior.

Second, *effectance* represents the motivation to interact effectively with the environment, and with non-human agents when applied to anthropomorphism. Its goal is to enhance one's ability to understand complex stimuli in the present and predict future behavior. However, if uncertainty regarding the behavior of a real or presumed non-human agent is activated by observing the agent's or non-human agent's behavior, anthropomorphism should be heightened. Firstly, uncertainty may arise when one is unfamiliar with the agent. Additionally, it can arise when the behavior of an agent appears unpredictable or inconsistent with one's expectations. When the actions of a non-human agent do not align with what one expects, one could be inclined to attribute human-like characteristics to try and make sense of the behavior. Lastly, when the underlying causal mechanisms are unknown or can not be directly observed uncertainty could arise. When lacking an understanding of the reasons behind a non-human agent's behavior, anthropomorphic thinking may be used to provide a sense of explanation. Besides uncertainty, anthropomorphism is likely to be heightened when there are strong incentives to accurately understand or predict the behavior of a non-human agent. For example, agents that are perceived as threatening or able to influence one's well-being may elicit anthropomorphic behavior more than powerless agents. Additionally, agents who one is likely to interact with more in the future are anthropomorphized more.

Lastly, *sociality* defines the need to establish and maintain a sense of social connection with others, which need can be easily satisfied by non-human agents. Sociality enhances the accessibility of social clues, including human-like characteristics. Individuals who are motivated to seek social interactions and connections are more likely to see human-like attributes in non-human agents. When people are deprived of social connections, they become more attentive to these social clues. These people have a stronger tendency to anthropomorphize non-human agents as a way to try to fulfill this need. The work of Epley et al. [6] is particularly relevant to our research as it offers a robust three-way theoretical framework containing factors that influence the amount of anthropomorphic behavior elicited, providing crucial points that can be used to do measurements. By drawing upon their framework, we can utilize solid factors that might influence the degree of anthropomorphism in our research.

Research has also been conducted concerning personal connections and emotions in the context of technology. Various terms can be used to describe a personal relationship with a piece of technology, like emotional attachment and affective quality [12]. Zhang [15] describes that technology has an affective quality if it can cause changes to a person's mood, emotions, and/or feelings. Emotional attachment can be described by looking at how some technologies change people's first impressions or engagement the more someone interacts with it. Sung [12] has shown this to be the case when researching human interaction with Roomba, an automatic vacuum cleaner. He found that users anthropomorphize this robot and were describing it with lifelike properties. Participants also felt that Roomba had intentions, feelings, and unique characteristics of its own. Besides emotional attachment to technology, Terzis et al. [14] have shown that emotional feedback affects the behavioral intent to use chatbots, in particular computer-based assessment. The above-mentioned works are relevant to our study as they explore the affective quality and emotional attachment to technology, as well as research the impact of emotional feedback and the user's behavioral intent to use such systems, indicating the importance of emotional factors in user interactions.

A different study that researched anthropomorphic language in the context of social media is one by Carter et al. [2]. First, they found that language concerning robots often exhibited human-like traits, therefore robots were mistaken for humans. Continuing on this they found that tweets about animals often contain human-like language as well, and caused significant mislabeling. This human-directed language contains modal constructions that use anthropomorphic explanations for behavior and three human-oriented grammatical features; pronouns, infinite verb forms, and "so" as a connector [11]. When individuals use pronouns like "he", "she" or "who" when referring to animals, it suggests that they view these entities as being more animate and possessing similar attributes to humans. However, the use of pronouns like "it", "which", or "that" suggest a more impersonal and objectifying perspective, indicating a lower level of anthropomorphism. These works are of great significance to our study as they give us a foundation for measuring the amount of anthropomorphic behavior in language.

When speaking about anthropomorphism in this paper, we refer to Salles et al. [10], which describes it as being "generally defined as the attribution of distinctively human-like feelings, mental states, and behavioral characteristics to inanimate objects, animals, and in general to natural phenomena and supernatural entities", together with the three-factors that make up this phenomenon given by the study of Epley et al. [6].

3 METHODOLOGIES

3.1 Data Collection

3.1.1 Crawler. The Twitter platform was utilized to create a dataset for this analysis. As the Twitter API had seen some changes recently, the decision was made to gather data through a web crawler. This tool spawned a web driver instance, which automatically navigated to different tweets. It went through all available replies, extracted the relevant information from each, and merged it into a single file. By gathering replies from different tweets that are of interest, a customizable dataset was still put together quickly.

3.1.2 Accounts. We established some predefined labels, to which we can link some accounts.

Ghostwriter. Accounts labeled as `ghostwriter` contained tweets written by humans, acting as if they come from objects.

@NASAPersevere - NASA's Perseverance Mars Rover NASA's Twitter account, imposing as the robot currently driving around on Mars.

Mechanic. Accounts labeled as `mechanic` contained tweets focused on providing numerical data in a systematic format.

@ReplyGPT This account's main interactions are within the replies to other tweets. Users may mention this account, which will then in turn use OpenAI's ChatGPT API. Though, it also does a daily post of its statistics and gives an insight into how many people it replied to. The format for this post starts by saying how many tweets it replied to that day, after which it thanks its users. While the precise phrasing might slightly differ, the tweets remain very similar.

Random. Accounts labeled as `random` contained tweets with no logical way of predicting the next tweet.

@ExplainThisBob A Twitter account, similar to *@ReplyGPT*, which can be tagged under tweets and generates a simplified explanation of that tweet. However, this account also tweets things on its own, with seemingly no connection between them.

When choosing these accounts for the analysis, we tried to get some variation in our dataset for the factors that are of interest, to be able to do an effective analysis.

Similarity. In terms of similarity, we tried to capture a wide variety of values. By choosing an account like ExplainItBob, we tried to saturate the lower spectrum of similarity values, as there does not seem to be any apparent relationship between the tweets. Furthermore, by choosing NASAPersevere we aimed at capturing a moderately high range of similarity values, as the tweets have the same head topic, exploring Mars, but should talk about different things it encounters. Lastly, we tried to populate the higher end of the scale by choosing ReplyGPT, which contains tweets that are very similar to each other in terms of structure and topic.

Sentiment. ExplainItBob seemed like an appropriate account with neutral sentiment, as the tweets contained very general language, were short, and did not show any real emotion. With NASAPersevere we tried to capture tweets that showed some level of positivity, as we expected the rover to positively report back on its findings. By choosing ReplyGPT we were aiming for the upper spectrum of sentiment values, as each message was very positive and contained a thank you to its followers.

When we will be analyzing the interactions with these different accounts, the labels are not of importance for the first two research questions, as we will look at factors across the entire dataset. However, the third research question analyses the influence of the author of an account. In this case, there is a distinction made between accounts. ReplyGPT and ExplainItBob will be grouped, as these both produce messages that are computer generated, in comparison to the human-written ones of NASAPersevere. When talking about *author* in this paper, we refer to this distinction.

It should be mentioned that there could have been additional accounts to make our dataset more inclusive. In terms of similarity in tweets, which will also be evident from subsubsection 4.2.1, ReplyGPT (generated) and NASAPersevere (ghostwritten) score relatively the same. Therefore, a ghostwritten random account would have been a good addition to the dataset, to have a more in-depth analysis. For sentiment, a more negative account could have been a good addition to saturate the lower end of sentiment values.

3.2 Data Processing

3.2.1 Cleaning. Before the dataset was analyzed, cleaning was done. As we saved all replies we could find, we encountered some that are not of interest. Examples of this were tweets that only contain an image, video, or GIF. Other tweets contained encoding errors or issues with special characters. These deficiencies will influence linguistic analysis, so they were discarded.

Additionally, we were only interested in English-written tweets. We used the `langdetect` [4] package that can detect a particular language. It outputs a certainty, in the range of `[0..1]`, together with a language. A threshold of 0.95 was set that needed to be achieved by the language detector, to ensure the quality of our dataset.

3.2.2 Natural Language Processing. After the dataset was cleaned several more processing steps were taken. The sentence was first converted to lowercase, after which the text was cleaned by removing unnecessary symbols. The content was then tokenized to break it down into simpler units. The remaining tokenized words were stemmed to return them to their base or root form. For this last step, the Porter stemmer from the NLTK [13] module was used.

3.2.3 Independent Samples Test. As we selected arbitrary accounts that fit into one of the pre-defined strategies, we needed to ensure that they were appropriate for the analysis. If the accounts turned out to be from the same sample, any derived measures would be rendered irrelevant, as we failed to capture the intended populations. Therefore, an independent sample test had to be performed on the dataset to verify whether the samples taken from the accounts were distributed differently. A Kruskal Wallis test was used on the different accounts with regard to similarity and sentiment. This resulted in pairwise comparisons of each account and showed whether the independent variables are different for each category.

3.3 Measuring Tools

3.3.1 Use of pronouns. After the data was collected, cleaned, and processed, we looked at what factors across this data signify anthropomorphic behavior. As described in section 2, one factor that could be linguistically analyzed was to see to what extent pronouns were being used in replies to certain tweets. A dictionary was constructed containing pronouns, which was used to count the average number of pronouns used per tweet.

3.3.2 Description of pronoun dictionary. To determine which pronouns to measure within replies we make use of an extended hierarchy presented by Foley [8].

(1) speaker/addressee > 3rd person pronoun > human proper noun > human common noun > other animate > inanimate

The hierarchy distinguishes categories that are closely related to humans and when shifting further right ones that represent a lower degree of anthropomorphism. This clear distinction allows for exact measurements of several grammatical concepts. We determined the first three to be the most appropriate to measure the level of anthropomorphism.

The dictionary used mainly consisted of the pronouns mentioned in Caulfield’s online article [3], with the addition of some contractions and abbreviations commonly used in informal communication, especially on social media. Examples of these contractions are "I’m", "I’ll", and "we’ll", where the first-person pronouns are not written out fully. Additionally, words like "you" are often abbreviated to "u" in short-form message contexts.

When looking into different accounts to be analyzed, we found that users often addressed accounts by their proper names. For example, ExplainItBob received a lot of comments along the lines of "How are you doing Bob?". Although the work from Sealey et al. [11] only touches the topic of pronouns, the extended hierarchy proposed by Foley [8] does include human proper nouns, therefore we opted to also count the use of human proper nouns when doing our measurements.

Pronoun Type	Pronouns	Additional Pronouns
First Person	’i’, ’me’, ’my’, ’mine’, ’myself’, ’we’, ’us’, ’ours’, ’ourselves’	’im’, ’ill’, ’well’, ’our’
Second Person	’you’, ’your’, ’yours’, ’yourself’, ’yourselves’	’youll’, ’youre’, ’u’, ’ull’, ’ur’, ’urs’
Third Person	’he’, ’him’, ’his’, ’himself’, ’she’, ’her’, ’hers’, ’herself’, ’they’, ’them’, ’their’, ’theirs’ ’themselves’	’hes’, ’hell’, ’shes’, ’shell’, ’theyll’

Table 1. An overview of the pronoun dictionary used to measure anthropomorphic language. Note that apostrophes are deprecated as these are removed during natural language processing.

3.4 Analysis

3.4.1 RQ 1: To what extent does similarity between a non-human agent’s messages influence the number of pronouns used by consumers? As can be read in section 2, effectance represents the motivation to interact effectively with the environment, but uncertainty may arise when an agent’s or non-human agent’s behavior is observed. One factor for this is unpredictability, and anthropomorphism should be heightened when an agent appears unpredictable or inconsistent. To measure the effect of this, similarity scores were calculated for each tweet, using the `spaCy` [7] Python package. Each tweet from one account was compared to the other tweets from that same author, resulting in a list of similarity values. The average of this list was taken, which gave a score that indicated how similar that tweet was to the other tweets from that account, giving an insight into how unpredictable their tweets were. To gauge the extent of this effect, the usage of pronouns was measured as described in subsubsection 3.3.1. The pairs of similarity scores with the percentage of pronouns from all accounts were merged into a single dataset, after which Kendall's rank correlation coefficient was calculated.

3.4.2 RQ 2: How does the sentiment of a non-human agent's messages correlate with the number of pronouns used by consumers? In addition to uncertainty, anthropomorphic behavior is more likely to be elicited when there are strong incentives to accurately understand or predict the behavior of a non-human agent. An example of this is when one perceives it as threatening or influencing someone well being. On top of this, Terzis et al. [14] have shown that emotional feedback, communication with an emotional non-human agent, affects the behavioral intent to use such an agent. To see whether there was a correlation between the sentiment of tweets and anthropomorphic behavior, a similar method to SRQ 1 in subsubsection 3.4.1 was used. The VADER NLTK [13] module was used to measure the positivity, neutrality, and negativity of tweets, which resulted in a compound score. Again, the pairs of sentiment score and percentage of pronouns from all accounts were combined, from which Kendall's rank correlation coefficient was computed.

3.4.3 RQ 3: To what extent does the author affect the use of pronouns in consumer replies? Lastly, we separated the tweets by author and analyzed whether one is more successful than another. The ExplainItBob and ReplyGPT accounts were combined into one category, as explained in subsubsection 3.1.2. To measure the effect of the author on the use of pronouns, we calculated a point-biserial correlation coefficient, specifically a Pearson bivariate correlation coefficient, as we deal with continuous data.

4 RESULTS

4.1 Description of the dataset

4.1.1 Cleaning. The dataset for the analysis consisted of several tweets for every strategy from each account mentioned in subsubsection 3.1.2. For each tweet, all available replies were scraped and incorporated into the dataset. This resulted in a total of 112 tweets, which contained 5381 replies, which after cleaning came down to 3438 usable replies, with approximately 1000 replies for each account.

Strategy	Handle	Tweets	Scraped Replies	Usable Replies
Random	@ExplainThisBob	16	2183	1090
Ghostwriter	@NASAPersevere	25	1737	1238
Mechanic	@ReplyGPT	71	1461	998

Table 2. A comprehensive overview of each strategy, along with the corresponding Twitter account and the respective counts.

As mentioned in subsubsection 3.1.2, some of the accounts analyzed supply users with the option to mention them under tweets to generate a reply. This results in these accounts often getting mentioned within their own replies, which resulted in a substantial amount of replies scraped only containing the mentioning of such accounts. The decision was made to not filter out these replies, as they arguably signify non-anthropomorphic behavior. When a user deliberately mentions a bot account to generate a reply, they are aware of the mechanisms behind it and therefore do not see it as a social entity. As this is linked to the strategy of an account, it should not be disregarded.

In the end, the average word count per reply was found to be 6.6 words, with a standard deviation of 4.7, ranging from a length between a minimum of 1 and a maximum of 56.

4.1.2 Natural Language Processing. Each entry of the dataset was processed with NLP techniques. This resulted in a certain degree of standardization across all entries, making the analysis less subject to small indifferences.

4.1.3 Independent Variables. A Kruskal Wallis test has been performed on the different factors of interest. With this, we can illustrate how the three different samples from different accounts taken overlap or are disjoint from each other.

Similarity. When observing the scatterplot in Figure 1, we notice that the similarity between the accounts seems different. ExplainItBob clearly shows different similarity values, whereas the other accounts seem close, but ever so slightly different. This is confirmed by the Kruskall-Wallis test across the accounts. The similarity turned out to be significantly different, with a Kruskal Wallis test statistic of 63.833, with a significance value of < 0.001, at a 1% significance level. The pairwise comparisons showed that all accounts were significantly different from each other in terms of similarity. The mean similarity scores for ExplainItBob, NASAPersevere, and ReplyGPT were $0.48(SD = 0.10)$, $0.82(SD = 0.02)$, and $0.86(SD = 0.04)$ respectively.

Sentiment. Looking at the scatterplot in Figure 2, we can see tweets from ExplainItBob and NASAPersevere have similar values of sentiment, whereas ReplyGPT seems to score high consistently. The Kruskal-Wallis test confirmed this, resulting in a value of $H = 52.213$ with a significance value of < 0.001, at a 1% significance level. From the pairwise comparisons, only ReplyGPT turned out to be independent of the other accounts at a significance value of < 0.001, significant at 1%. The comparison between ExplainThisBob and NASAPersevere received a significance value of 0.377, therefore not showing a significant difference in distribution at a significance level of 1%. The mean sentiment scores for ExplainItBob, NASAPersevere, and ReplyGPT were $0.26(SD = 0.32)$, $0.38(SD = 0.34)$, and $0.78(SD = 0.11)$ respectively.

From this, we can conclude and establish, when talking about similarity and sentiment in the rest of the paper, we have significant differences in these factors between accounts.

4.1.4 Dependent Variables.

Use of pronouns. To display the usage of pronouns across all replies, the mean was taken across the number of pronouns used and normalized by dividing the amount by the number of words for each reply. This resulted in the first- and second-person pronoun categories exhibiting a mean of 0.03 ($SD = 0.06$) and 0.03 ($SD = 0.07$) respectively, suggesting a comparable usage level. In contrast, the

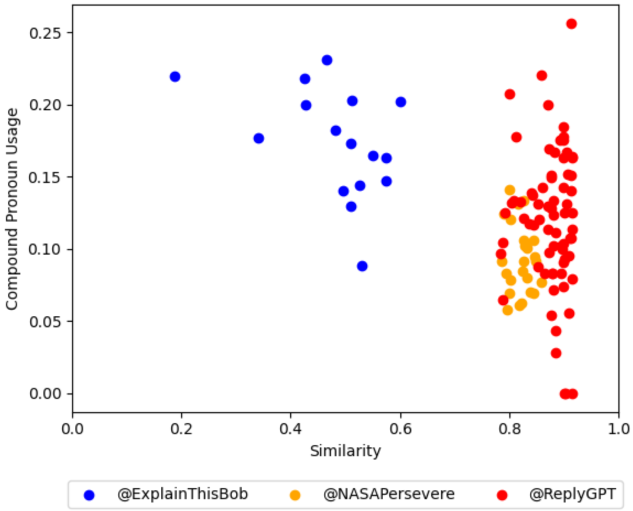

Fig. 1. Scatterplot, per author, showing the relation between the similarity of a tweet and the use of pronouns in the replies of that tweet.

third-person pronoun category displays a lower mean of 0.02 with a standard deviation of 0.05, suggesting an overall lower usage and relatively lower variability. The names category shows a mean of 0.02 with a standard deviation of 0.07, indicating an overall moderate usage level. A single score was comprised of all pronoun types and the mentioning of names.

4.2 Findings

4.2.1 RQ 1: To what extent does similarity between a non-human agent's messages influence the number of pronouns used by consumers? The similarity across the entire dataset had a mean of 0.808 with a standard deviation of 0.143. These values were in line with the expectation, as only one of the accounts analyzed seemed to have no coherent topic between tweets, while the others stayed close to the same one.

This distribution can be observed in Figure 1, where the majority of data points are skewed towards the right-hand side. Additionally, there appears to not be any discernible correlation between the similarity of tweets and the use of pronouns. We tried to further back this up with the result of Kendall's tau-b correlation test. It resulted in a τ_b correlation coefficient of 0.121, with a significance value of 0.060. At a significance level of 5%, we are not able to report any significant result from our dataset. Therefore, additional experiments may be performed to show the possible correlation between variables.

4.2.2 RQ 2: How does the sentiment of a non-human agent's messages correlate with the number of pronouns used by consumers? Across the entire dataset, the mean sentiment score was 0.618 with a standard deviation of 0.308. The observed distribution of data points along the x-axis was in line with our expectations. The tweets from ReplyGPT were all very positive, as they were thanking their followers in a very systematic way, resulting in a very positive score across all tweets, which can be seen in Figure 2.

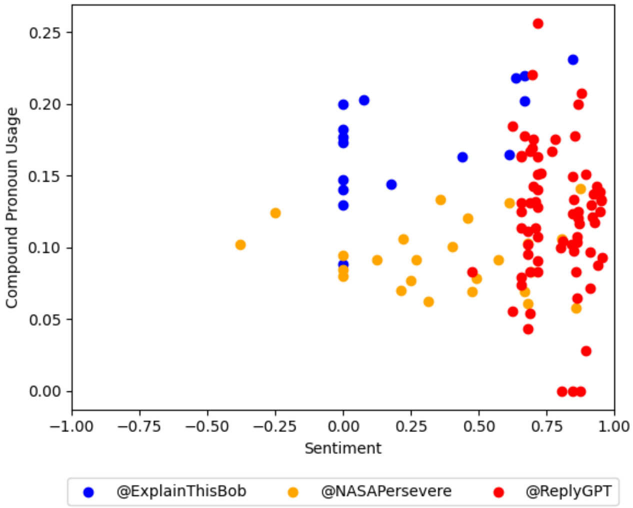

Fig. 2. Scatterplot, per author, showing the relation between the sentiment of a tweet and the use of pronouns in the replies of that tweet.

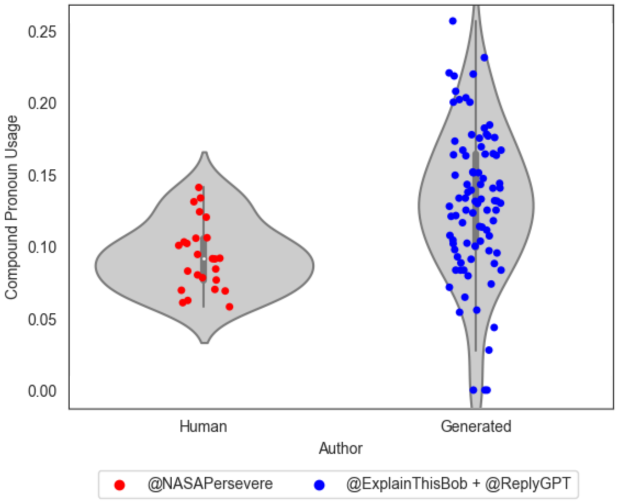

Fig. 3. Violin plot overlayed with a jittered strip plot, per author, showing the density and distribution of the pronoun usage across the category.

From Figure 2, there does not appear to be a significant trend visible. To try and confirm this, Kendall's correlation coefficient rank was computed on sentiment and the use of pronouns, resulting in $\tau_b = 0.003$ and a significance value of 0.958. Therefore, from this dataset, we were not able to reach any significant result at a significance level of 5%, and further experimentation should be done.

4.2.3 RQ 3: To what extent does the author affect the use of pronouns in consumer replies? After combining the data per author we were left with our dataset divided into two different categories: human-written and computer-generated messages. The violin plot overlayed with a jittered strip plot in Figure 3 shows the density of the pronoun usage across the different tweets in that category. Visually from this, it can be seen that the generated tweets have a wider range of pronoun usage in replies, but also on average show a higher use of pronouns.

The Pearson bivariate correlation coefficient supports this assumption. With generated messages being categorized as 0 and human-written messages categorized as 1, the correlation test resulted in a correlation coefficient of -0.316, with a significance value of < 0.001, at a 5% significance level. This suggests that there is a moderate negative correlation between human-written messages and the use of pronouns in replies by consumers.

5 DISCUSSION

Our results showed that we were not able to establish any significant correlation between the similarity of messages and the amount of anthropomorphic behavior elicited (in terms of pronoun usage in replies), which is not in line with theory [6]. A factor that could have influenced this result is the time frames the scraped tweets were chosen from. We decided to scrape consecutive tweets until a suitable amount of usable replies was reached, instead of cherry-picking specific tweets to analyze. This might have resulted in similarity scores being closer than anticipated, especially for the NASAPersevere account. The tweets that were covered in the time frame were centered around a certain mission that involved dropping 10 sample tubes on Mars, making the topic and word choices very similar. However, an argument can also be made that this is part of the strategy employed by the ghostwriters and that a different account should have been selected.

Additionally, we were not able to measure any significant correlation between the sentiment of tweets and the use of pronouns in the replies, contradicting existing literature [14]. One possible explanation for this could be that the users interacting with these accounts only encounter them occasionally, as they are dependent on the Twitter algorithm if they do not follow the accounts. Therefore, they might not have developed a sense of familiarity or connection with the accounts, potentially reducing the impact of any emotional effects.

Another more general explanation for these results could be linked to elicited agent knowledge from the three-factor theory [6], as described in section 2. As briefly mentioned in results, users interacting with accounts that generate a reply when mentioned, are generally more aware of the mechanism behind such an account. Therefore, they might have an easier time overcoming default assumptions about these non-human agents, showing a lower amount of anthropomorphic behavior compared to regular users who come in contact with these accounts.

We did observe a significant correlation between the use of pronouns and the author of an account. Computer-generated messages showed a higher level of pronoun usage in replies, which contradicted our hypothesis. We expected ghostwritten content to show higher levels of anthropomorphism, as we predicted that human authors can mimic and display human-like attributes more effectively. One possible explanation for this unexpected result could be that computer-generated messages are designed to imitate human-like language, enhance user engagement, and create a conversational experience, whereas the ghostwritten account chosen was more focused on sharing knowledge.

Lastly, sometime after the tweets were scraped and the analysis was performed, the ExplainThisBob account was suspended from Twitter, due to allegations of being a “crypto scam account" [9]. The presence of a suspended account in our dataset possibly should be taken into account when interpreting these results. The reason for suspension might raise questions regarding the reliability and trustworthiness of the content and interactions of an account. Additionally, the reproducibility of this study is affected, as it hinders the possibility to reproduce these exact results, due to the content and interactions with this account are no longer available.

5.1 Limitations

The dataset that was created for this analysis was limited due to various reasons. Firstly, a drawback of our crawler was that Twitter only serves users with a certain subset of all replies under a tweet. The threshold for this lies around 200-250 replies, after which no new replies will be loaded, giving us less control over replies, resulting in ones that are not interesting to us, e.g. images, gifs, or video replies. Additionally, ReplyGPT did not receive nearly as many replies as the other accounts analyzed. Therefore, the number of tweets scraped for each strategy to get a similar amount of usable replies is out of proportion.

Furthermore, our findings contradicted existing literature, for which there could be multiple explanations. Firstly, the method for measuring anthropomorphic behavior in linguistics may have inherent limitations. It remains a difficult problem to measure a psychological phenomenon using linguistic analysis, especially in the context of social media, where there is an overall inclination towards concise messages. The way of measuring sentiment within tweets also caused some limitations. VADER is not the most robust way to measure sentiment in a tweet, and as shown in Figure 2, tends to give neutral scores, resulting in a higher probability of getting tied ranks when computing Kendall's rank correlation coefficient. Additionally, it only produces values on a single axis, therefore only showing whether a piece of text is considered to be positive or negative.

5.2 Future Works

With the continuous improvement of tools like ChatGPT and the increase in the use of social media and other platforms in our lives, there are many different opportunities to expand upon this work.

In the case of measuring anthropomorphism, one could explore a better way of measuring anthropomorphism within language. The one we used is appropriate, but there are underlying or hidden factors that have an impact on the results. A more robust and broader technique might allow for more accurate results.

One could also explore different factors that could influence the amount of anthropomorphic behavior elicited in consumers. One factor in particular that would be interesting in examining further, is the one of giving a non-human agent a short and easy name. When looking at different accounts that were of interest, it became apparent that a lot of replies use the author's name, e.g. ExplainItBob was called “bob" a lot. A very easy-to-understand name might make an agent feel more personal, in contrast to something like the “NASA Mars Perseverance Rover". Continuing on this, ExplainItBob and ReplyGPT both had an image of characters with some human characteristics as their profile picture, whereas NASAPersevere had one of their rover. Again, the notion of visual appearance might play a significant role in the way users anthropomorphize accounts. Therefore, an experiment could be performed to examine the effect appearance and characteristics of an account.

As briefly mentioned in subsubsection 3.1.2, the addition of other accounts could be something to improve upon. A 2-by-2 interaction effect could for example be studied if we had analyzed a ghostwritten random account as well. This way we would have two random and two similar accounts, wherein each category one is generated and one is ghostwritten. This allows for a more in-depth analysis and might show different results than the ones we were able to produce.

More observation is also needed in seeing how the interaction with these accounts compares to real humans. Are the similarities to be found? Do these accounts show significantly higher pronoun usage or other measures? This could either be done through another analysis or an experiment.

Lastly, it is important to note possible ethical concerns with the creation of technology like this, especially with the rise of AI-generated content. Nowadays it can be hard to distinguish human-written content from computer-generated content, and the question rises whether it is considered ethical to not disclose this information. Users might build unsustainable personal relationships with technology, even though this is not the intended use. The explicit mention of the use of generated content might be a way of mitigating this concern, as being transparent could de-incentivize the creation of such relationships. However, an in-depth literature review or experiment should be performed to explore when and how users build relationships with technology and possible ways of avoiding this. Only after this is established, one could explore the development of such technology.

6 CONCLUSION

In conclusion, our findings show that the author, human-written or computer generated, influences the amount of anthropomorphic behavior elicited. Generated messages seem to show higher levels of pronoun usage in their replies, however, the factors analyzed did not seem to have a substantial impact. We were not able to measure any significant correlation between dissimilar tweets and a heightened amount of anthropomorphism within our dataset. Additionally, our results showed sentiment to not be a significant factor either. Therefore, we were not able to reach our goal, as we are still unsure what effect the factors analyzed have on anthropomorphism. This means we are still not sure how a social entity should communicate and are unsure whether we can bridge the gap between producers and consumers. We highlight the need for further research to explore different ways of measuring anthropomorphism and different factors that influence the amount of anthropomorphic behavior. Lastly, we emphasize the potential drawbacks such outcomes might have on consumers, which should be taken into account when expanding upon this or similar work.

REFERENCES

[1] L.R. Caporael. [n. d.]. Anthropomorphism and mechanomorphism: Two faces of the human machine. 2, 3 ([n. d.]), 215–234. https://doi.org/10.1016/0747-5632(86)90004-X

[2] Elizabeth J. Carter, Samantha Reig, Xiang Zhi Tan, Gierad Laput, Stephanie Rosenthal, and Aaron Steinfeld. [n. d.]. Death of a Robot: Social Media Reactions and Language Usage when a Robot Stops Operating. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, 2020-03-09) *(HRI '20)*. Association for Computing Machinery, 589–597. https://doi.org/10.1145/3319502.3374794

[3] Jack Caulfield. [n. d.]. *What Is a Pronoun? | Definition, Types & Examples.* https://www.scribbr.com/nouns-and-pronouns/pronouns/

[4] Michal Mimino Danilak. [n. d.]. *langdetect: Language detection library ported from Google's language-detection.* https://github.com/Mimino666/langdetect

[5] Rachel Duffy, Andrew Fearne, and Victoria Healing. [n. d.]. Reconnection in the UK food chain: Bridging the communication gap between food producers and consumers. 107, 1 ([n. d.]), 17–33. https://doi.org/10.1108/00070700510573177 Publisher: Emerald Group Publishing Limited.

[6] Nicholas Epley, Adam Waytz, and John T. Cacioppo. [n. d.]. On seeing human: A three-factor theory of anthropomorphism. 114 ([n. d.]), 864–886. https://doi.org/10.1037/0033-295X.114.4.864 Place: US Publisher: American Psychological Association.

[7] Explosion. [n. d.]. *spacy: Industrial-strength Natural Language Processing (NLP) in Python.* https://spacy.io

[8] W. A. Foley and Robert D. Van Valin Jr. [n. d.]. *Functional syntax and universal grammar.* Cambridge University Press. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_145877

[9] Decrypt / Tim Hakki. [n. d.]. *'Explain This Bob' AI Twitter Bot Suspended After Elon Musk Calls it 'Scam Crypto Account'.* https://decrypt.co/145206/explain-this-bob-ai-twitter-bot-suspended-after-elon-musk-calls-it-scam-crypto-account Section: News.

[10] Arleen Salles, Kathinka Evers, and Michele Farisco. [n. d.]. Anthropomorphism in AI. 11, 2 ([n. d.]), 88–95. https://doi.org/10.1080/21507740.2020.1740350 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21507740.2020.1740350.

[11] Alison Sealey and Lee Oakley. [n. d.]. Anthropomorphic grammar? Some linguistic patterns in the wildlife documentary series Life. 33 ([n. d.]), 399–420. https://doi.org/10.1515/text-2013-0017

[12] Ja-Young Sung, Lan Guo, Rebecca E. Grinter, and Henrik I. Christensen. [n. d.]. “My Roomba Is Rambo”: Intimate Home Appliances. In *UbiComp 2007: Ubiquitous Computing* (Berlin, Heidelberg, 2007) *(Lecture Notes in Computer Science)*, John Krumm, Gregory D. Abowd, Aruna Seneviratne, and Thomas Strang (Eds.). Springer, 145–162. https://doi.org/10.1007/978-3-540-74853-3_9

[13] NLTK Team. [n. d.]. *nltk: Natural Language Toolkit.* https://www.nltk.org/

[14] Vasileios Terzis, Christos N. Moridis, and Anastasios A. Economides. [n. d.]. The effect of emotional feedback on behavioral intention to use computer based assessment. 59, 2 ([n. d.]), 710–721. https://doi.org/10.1016/j.compedu.2012.03.003

[15] Ping Zhang and Na Li. [n. d.]. The importance of affective quality. 48 ([n. d.]), 105–108. https://doi.org/10.1145/1081992.1081997