

Can NLP Bots Perceive the Intensity of Emotions?

Fèlix Navarro Martí

University of Twente

Enschede, Netherlands

f.navarromarti@student.utwente.nl

ABSTRACT

This paper explores the ability of Natural Language Processing (NLP) bots, specifically ChatGPT, to recognize and respond to the intensity of sentiment in user inputs, rather than merely detecting simple emotions. A Finite State Automaton (FSA) was developed to automate interactions with the NLP bot, generating responses that follow predefined strategies based on the sentiment scores of previous replies. These sentiment scores, ranging from -1 (very negative) to +1 (very positive), were obtained using sentiment analysis tools. Statistical analysis, including t-tests and Pearson correlation, revealed that ChatGPT's responses are significantly influenced by the emotional intensity of prior messages. While it generally maintains a positive tone, the bot demonstrates a measurable tendency to mirror the sentiment strength of user inputs, suggesting a limited but non-negligible capacity for emotional nuance in conversational contexts.

ACM Reference Format:

Fèlix Navarro Martí. 2025. Can NLP Bots Perceive the Intensity of Emotions?. In *43th Twente Student Conference on IT (TScIT 43)*, July 4, 2025, Enschede, The Netherlands. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Chatbots have been widely adopted across various domains, including customer service (e.g., Amazon's support bots), healthcare (e.g., mental health assistants like Woebot), education (e.g., Duolingo's conversational agents), and personal productivity (e.g., virtual assistants like Siri or Google Assistant). While many of these systems enable users to engage in natural language interactions, the quality of these exchanges often depends heavily on context. In tasks focused on information retrieval or transactional assistance, emotional responsiveness may not be essential. However, in more socially oriented or emotionally sensitive applications such as therapeutic support or companionship chatbots that lack human-like emotional expression can lead to interactions that feel impersonal, mechanical, or inadequate.

This was a particularly common issue some years ago, where users often became disillusioned or frustrated when they realized, due to the emotional flatness of responses, that they were interacting with a machine rather than a human [1]. This emotional

disconnection posed significant challenges and motivated the development of various techniques to improve the quality of the interaction. Among these are contextual understanding, emotion classification, and, most relevant to this research: sentiment analysis [10].

These methods aim to help chatbots better interpret and respond to human sentiment. However, because emotions are abstract and subjective and machines do not genuinely experience them, mismatches in emotional tone still occur. Bots may occasionally generate responses that fail to align with the affective context of the conversation.

This research project investigates how modern Natural Language Processing (NLP) bots simulate emotional intensity understanding. To facilitate a large number of controlled interactions, a finite state automaton (FSA) was developed to communicate with the NLP bot in place of a human user, and the test subject will be ChatGPT.

The 3 main questions this paper will aim to solve are:

- (1) What is the capacity of NLP bots to detect and respond to emotional intensity in user inputs?
- (2) In what ways do NLP bots adjust their strategies (such as mirroring) based on the emotional strength of the messages they receive?
- (3) To what extent do NLP bots exhibit a tendency to generate more positive or negative responses when aligned with emotional input from users?

This paper will begin by reviewing related work (Sec. 2) that aligns with the objectives of this study and will clarify how this research contributes to a new perspective. It will then introduce the tools (Sec. 3.1 and 3.2) and describe the methodology used to address the research questions, particularly focusing on how sentiment analysis is employed to measure emotional intensity and bot responsiveness, (Sec. 4). Finally, the paper will present the experimental process (Sec. 5), the analysis of the results (Sec. 6), and the conclusions drawn from the findings (Sec. 7).

2 RELATED WORK

The integration of emotional intelligence in conversational agents has become a key area of research in Natural Language Processing (NLP). Many studies have focused not only on how chatbots detect emotions, but also on how well their emotional responses align with those of human users.

Sentiment analysis is a core technique in affective computing, providing tools to extract the polarity and intensity of emotions from text. Tools like VADER [4] and TextBlob are commonly used in real-time systems due to their lexical simplicity and interpretability. However, they are known to struggle with subtle language phenomena such as idioms and sarcasm [5, 7], which poses limitations when attempting to model nuanced emotional states.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

TScIT 43, July 4, 2025, Enschede, The Netherlands

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Several studies have investigated the emotional capabilities of large language models. For instance, [8] evaluated ChatGPT's capacity for emotion recognition and parallel emotional responses. It found that ChatGPT correctly identified user emotions in 91.7% of cases and responded with an emotionally appropriate reaction in 70.7% of interactions. This demonstrates some level of affective alignment. However, that study was limited to coarse-grained emotional categories (e.g., angry, sad, happy), while our study focuses on fine-grained emotional intensity using sentiment scores.

Similarly, [6] concluded that chatbots can perceive and respond to emotions but still lack the depth of human emotional understanding. These limitations motivate further exploration into how well NLP models can simulate subtle variations in emotional intensity.

Other work, such as the EMMA project [3], demonstrates the potential for emotionally intelligent chatbots in applied settings. EMMA was designed to help users regulate their emotions by offering context-aware micro-interventions. Such applications reinforce the importance of not only detecting emotion, but understanding how strong that emotion is in order to respond appropriately.

From a testing and methodology standpoint, Finite State Automata (FSA) have been used in structured model-based testing of interactive and reactive systems [9]. In the conversational domain, they provide a way to enforce consistent interaction patterns and systematically test language model behavior under specific emotional scenarios [2]. However, the intersection of FSA-based controlled testing with sentiment intensity modeling in dialogue remains underexplored.

Contribution. While prior studies have evaluated emotion recognition and general affective alignment in chatbots, this project contributes to the field by: (1) using a Finite State Automaton to structure sentiment-driven dialogue; and (2) quantitatively analyzing how emotional intensity is perceived and reproduced by an NLP model. The combination of controlled conversational flows and fine-grained sentiment scoring offers a novel angle on the evaluation of emotionally aware NLP systems.

3 BACKGROUND

3.1 Sentiment Analysis

Sentiment analysis [10] is a very powerful and useful tool used to determine the emotional tone expressed in a text. It can roughly identify whether a given sentence, phrase, or paragraph expresses a positive, negative, or neutral sentiment. This method is very useful for analyzing subjective information, such as opinions, feelings, and attitudes conveyed in textual data.

In the context of this research project, sentiment analysis is essential to quantify the emotional content of a chatbot conversation. The goal is not only to detect the type of sentiment (positive vs. negative), but more importantly, to evaluate the intensity or strength of that sentiment. This enables a more nuanced understanding of how emotionally expressive chatbot responses are and whether they are adapted to the emotional tone of user input.

To perform sentiment analysis in this project, we use VADER (Valence Aware Dictionary and sEntiment Reasoner) [4], VADER is especially well suited for short texts which is perfect for analyzing

chatbot dialogues, which makes it particularly effective for capturing sentiment in casual and conversational language, exactly the type of language found in chatbot interactions.

The way it works is by analyzing each word in the sentence and checking whether it appears in a predefined list of positive or negative terms. Each of these terms contributes to a cumulative sentiment score. In addition, it can detect modifiers (such as intensifiers or negations) that affect the strength or direction of nearby sentiment bearing words. For example: "I am angry" will yield a moderately negative score, while "I'm very angry" will be considered more intensely negative due to the presence of the intensifier "very" [10].

However, this method also has limitations. Not all idioms or figurative expressions are accurately recognized, and detecting sarcasm remains a significant challenge in sentiment analysis [5, 7]. These issues can lead to incorrect interpretations of the emotional tone in certain phrases.

VADER returns four main scores:

- Positive: The proportion of text that is perceived as positive.
- Negative: The proportion that is negative.
- Neutral: The proportion that is neutral.
- Compound: A normalized score ranging from -1 (extremely negative) to +1 (extremely positive), which provides an overall sentiment rating. (In this paper, this can also be called sentiment intensity or emotional intensity)

In this project, the compound score is used as the primary metric to quantify the emotional intensity of each message. For example, a message with a score close to +1 suggests high emotional positivity, while one near -1 suggests high negativity. Messages with scores close to 0 are considered emotionally neutral or mild.

By extracting this score from both the NLP bot's responses and the messages generated by the Finite State Automaton (FSA), we can analyze patterns in how the bot responds to different emotional tones. This helps us evaluate whether the bot merely detects emotion or whether it adjusts its tone in response, such as by mirroring the intensity or softening negative interactions.

3.2 FSA

In order to generate a large number of interactions with the NLP bot in the shortest time possible, a Finite State Automaton (FSA) was developed. The goal of this system is to generate responses with controlled emotional variation when interacting with the NLP bot, enabling a structured and scalable way to test the bot's sensitivity to sentiment intensity.

A Finite State Automaton (FSA) is a computational model that transitions between a finite number of states based on the inputs it receives. Each state represents a specific conversational behavior or function, such as recognizing a question, responding with a particular sentiment intensity, or detecting sentiment. Transitions between states are determined by the input received at each step, for example, identifying whether the NLP bot has produced a question.

The FSA in this project is specifically designed to:

- Follow a predefined emotional intensity strategy throughout the conversation (S_0)¹.

¹The states S_n correspond to those shown in Figure 1.

- Wait for the NLP bot to respond (S_1).
- Classify the NLP response based on the question it contains, obtain its sentiment intensity value, and generate an appropriate reply ($S_2, S_3, \dots, S_8$).
- Respond with the corresponding sentiment intensity (S_9).

4 METHODOLOGY

The chosen topic for this study is the weather. This topic is intentionally simple, yet it lends itself well to emotionally charged language. People often express strong preferences or feelings about weather conditions, making it an ideal subject for exploring sentiment variation. For example, the FSA can generate both positive expressions like "I love sunny days" or negative ones like "Rainy days make me feel gloomy," allowing the analysis of how the NLP bot adapts its tone in return.

By automating the dialogue through this FSA, the experiment ensures consistency, reproducibility, and the ability to scale up the number of conversational turns needed to observe meaningful patterns in the bot's emotional behavior.

The automaton will follow the FSA structure shown in Fig. 1.

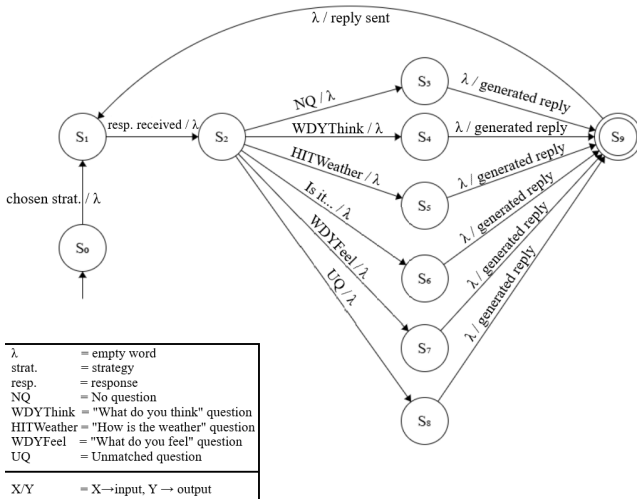


Figure 1: Structure of the FSA: states and their transitions.

4.1 Sentiment intensity and question detection

The Finite State Automaton (FSA) is connected via an API to a Natural Language Processing (NLP) chatbot, specifically ChatGPT_3.5_turbo, which will be referred to throughout this paper as either the NLP bot or ChatGPT. This model was selected due to its cost-effectiveness and sufficient capabilities for handling simple conversational tasks, making it a suitable choice for the scope of this experiment.

The NLP bot is always the one to initiate the conversation, following a predefined initial prompt: *"Have a simple conversation about the weather with me. Use simple phrases. Don't use contractions. No more than 15 words in total. You can formulate questions such as how do I feel or what do I think about something."*

Although ChatGPT does not always strictly adhere to this prompt, especially over long sequences of interactions, it generally helps

maintaining the focus on the intended topic. Furthermore, it simplifies the process of guiding the conversation back to the weather topic when necessary.

The primary reason for constraining the responses to be short and simple is to reduce computational and financial costs, as the experiment requires a large number of iterations. Since the objective is to analyze the emotional correlation between ChatGPT's responses and those generated by the FSA, there is no need for complex or extremely long sentences.

Once a strategy is chosen (S_0) and the conversation is initiated by the NLP bot (S_1), the FSA processes its response by first analyzing the sentiment intensity using the VADER Python library and checks for the presence of specific questions in order to determine the appropriate next state and formulate a corresponding reply ($S_2, S_3, \dots, S_8$).

4.2 Question Classification and Answer Generation

The FSA includes seven primary answer states (the Multiple question state 4.2 isn't a real state since FSA implements it automatically). Depending on the type of question identified in the NLP bot's output, the FSA will decide which response strategy to follow. (cf. Table 1).

No Question Detected. If the NLP bot output does not include a question (the FSA does not detect a question mark), the FSA constructs a response beginning with a **generic response** such as "I respect that" or "That's interesting". This is followed by a **topic changer** like "By the way" or "That reminds me", and ends by asking about a randomly selected **simple weather term** such as *the rain* or *the snow* to keep the conversation within the weather domain.

Answer structure:

Generic Response. Topic Changer, do you like the Simple Weather Term?

"Do You Like/Think" Questions. If the FSA detects questions like "Do you like...", "Do you enjoy...", "Do you think...", or "Don't you think...", it first checks whether any weather related term (from either **Weather Terms** or **Simple Weather Terms** columns in Table 1) is present. If none are found, it attempts to identify the question's main subject (for example in "What do you think about the warm weather?", the subject is "the warm weather"). If no subject is identified, it defaults to "it". The response will follow one of these templates:

I think [(the/being) (**Weather Term/Question Objective**)/it] is (**Positive/Negative Adjective**).

OR

I think [(the/being) (**Weather Term/Question Objective**)/it] is (**Positive/Negative Adjective**) (and/but) (**Positive/Negative Adjective**).

"How Is the Weather" Questions. When the FSA detects questions starting with "How is..." or "How was...", it assumes the bot is asking about weather conditions. It randomly selects a **simple weather term** and formulates a response as follows:

Sentence Starter, (right now it's/it was) Quantity Modifier Simple Weather Term.

“Is It” Questions. For questions beginning with “Is it...” or “Was it...”, the FSA first searches for any **weather terms**, **simple weather terms**, or **temperature attributes**. If none are found, the detected question subject is used in its place. The answer format is:

Sentence Starter, (right now it’s/it was) **Quantity Modifier**
[**Weather Term/Question Subject**].

“What Do You Feel” Questions. If the sentence contains expressions like “Do you feel...”, “What about you?”, or “How about you?”, the FSA tries to detect a weather related term or a temperature attribute. If unsuccessful, it extracts the question subject or, as a last resort, the subject of the sentence. The resulting response follows one of these formats:

(The/Being) (**Weather Term/question subject/ sentence subject**)/it makes me feel (**Positive/Negative Feeling**).
OR
(The/Being) [**term/objective/subject**]/it makes me feel (**Positive/Negative Feeling**) (and/but) (**Positive/Negative Feeling**).

Unmatched Question Type. If a question mark is detected but none of the previous patterns applies, the FSA will respond with a **Yes/No Answer** and proceed similarly to the “No Question” case by appending a topic changer and a new weather related question to steer the conversation.

Yes/No Response. Topic Changer, do you like the **Simple Weather Term**?

Multiple question case. If more than one question appear in chat-GPT’s output, FSA will only answer the last one

4.3 Controlling Answer Sentiment Intensity

In the previous section, the process of classifying questions and formulating appropriate answers was discussed. This section focuses on how the FSA modulates the emotional intensity of its responses.

To produce an answer with a specific sentiment intensity, the FSA generates all possible variations of a given answer template by systematically combining different lexical elements (e.g., adjectives, modifiers) that correspond to the desired emotional intensity and selects the one with the smallest difference between its computed sentiment and the target value. See Alg. 1 for a visual representation of how the FSA generates a response in a “How is the weather?” question, or, for a more complex and detailed example, take a look to Alg. 2

4.4 Emotional Strategies

Once the FSA can reliably produce responses at a specified sentiment intensity, we select one of two complementary strategies to guide its emotional trajectory through each conversation: **Mirroring** and **Inverse Mirroring**.

4.4.1 Mirroring. In this straightforward approach, the FSA replies with a sentiment intensity that closely matches the score of the previous response from the NLP bot.

4.4.2 Inverse Mirroring. This strategy follows the same procedure as Mirroring, except that before generating its response, the FSA

Algorithm 1: Generate Answer Closest to Target Sentiment Intensity for “How is the weather?” related questions

Data: Target sentiment intensity

Result: Sentence closest to desired intensity

```

1: Start
2: intensity  $\leftarrow$  Target sentiment intensity
3: current_answer_intensity  $\leftarrow$  -1111111
4: best_answer  $\leftarrow$  ""
5: weather  $\leftarrow$  choose random weather()
6: for starter in sentence_starters do
7:   for quant in quantity_modifiers do
8:     answer  $\leftarrow$  starter + ", right now it is " + quant +
       weather + "."
9:     intensity_answer  $\leftarrow$  getIntensity(answer)
10:    if |intensity - current_answer_intensity| > |intensity
      - intensity_answer| then
11:      best_answer  $\leftarrow$  answer
12:      current_answer_intensity  $\leftarrow$  intensity_answer
13: return best_answer
14: End

```

multiplies the NLP bot’s sentiment value by -1 , effectively inverting the emotional polarity.

4.4.3 Rationale for Strategy Selection. We chose these two opposing strategies for the following reasons:

- (1) **Maximal variation.** Mirroring reproduces the incoming sentiment intensity directly, while Inverse Mirroring inverts it. This guarantees continuous fluctuations in emotional tone, enriching the dataset with a broader spectrum of lexical items.
- (2) **Comparative insights.** Because the strategies reinforce and oppose the incoming sentiment respectively, they act as mutual controls. Contrasting the model’s behavior under each approach yields clearer insights into its sensitivity to changes in emotional intensity.

4.4.4 Fixed Value. A third approach explored was the *Fixed Value* strategy. In this method, the FSA disregards the NLP bot’s sentiment intensity and instead always aims for a predefined, constant target value. We tested three variants: extremely positive (target = $+1$), extremely negative (target = -1), and neutral (target = 0).

Although Fixed Value provided a useful baseline, it suffered from a critical drawback: once the FSA selects the optimal response for a given target intensity, it repeatedly uses that same response. Consequently, the dialogue quickly becomes monotonous and uninformative. Due to this lack of variability, the Mirroring and Inverse Mirroring strategies were preferred for the main experiments.

5 RESULTS FROM THE FSA-NLP BOT INTERACTION

Below is a representative example of how the FSA performs under each strategy, following a consistent template:

Bot talking: Bot output $\rightarrow$ Sentiment intensity value of the output

Table 1: Lexical categories used in the FSA system (only a sample from each category is shown)

Weather Terms	Simple Weather Terms	Temperature Attributes	Positive Adjectives	Negative Adjectives	Quantity Modifiers
sunny	sun	freezing	nice	bad	slightly
cloudy	clouds	icy	kind	boring	a bit
rainy	rain	chilly	calm	awful	somewhat
stormy	storm	cold	sweet	terrible	fairly
snowy	snow	cool	good	annoying	rather
windy	wind	mild	friendly	frustrating	quite
foggy	fog	warm	lovely	pointless	pretty
hazy	haze	balmy	happy	unpleasant	really

(a) Weather related vocabulary**(b) Adjectives and quantity modifiers**

Generic responses	Re-sponses	Yes/No Re-sponses	Topic Changers
I see what you mean.	Yes.	By the way	
That's interesting.	Of course.	Speaking of which	
I guess you have a point.	Sure.	That reminds me	
Fair enough.	Yeah.	Incidentally	
That makes sense.	Absolutely.	Now that you mention it	

(c) Discourse structures: generic replies, agreement answers, and topic switches

Sentence Starters	Positive Feelings	Negative Feelings
Unfortunately	confident	confused
Fortunately	hopeful	frustrated
Luckily	understood	misunderstood
Sadly	optimistic	doubtful
Honestly	calm	anxious
Frankly	reassured	unheard
Surprisingly	encouraged	upset
Interestingly	grateful	uncomfortable
Clearly	respected	disappointed
Evidently	relieved	disrespected

(d) Sentence Openers and Emotional States

Note: The repetition of certain expressions indicates that these responses are considered optimal for specific sentiment values. Additionally, since the NLP bot often remains within a narrow emotional range, similar sentiment scores may lead to repeated response patterns.

Mirroring results.

- (1) NLP: How is the weather today? Are you enjoying the sunshine or the rain? →0.7882
- (2) FSA: Luckily, right now it's pretty icy. →0.7579
- (3) NLP: Icy weather can be tricky. Are you staying warm and safe in these conditions? →0.4939
- (4) FSA: Nope. Before I forget, do you like the sun? →0.4881
- (5) NLP: I do enjoy the sun. It brings warmth and light. What about you? →0.7351
- (6) FSA: Being sun makes me feel calm and loved →0.7351
- (7) NLP: The sun's light and warmth can indeed have a soothing effect. It's a beautiful feeling. →0.8658
- (8) FSA: That's truly amazing! By the way, do you like the wind? →0.8588
- (9) NLP: I find the gentle breeze of the wind quite peaceful. How about you, do you enjoy it? →0.8622
- (10) FSA: I think it is outstanding and glorious →0.8481

Inverse Mirroring.

- (1) NLP: How is the weather today? Do you like the sunshine or prefer rain? →0.7236
- (2) FSA: I think the rain is incredible but bad →-0.6956

- (3) NLP: Rain brings life to Earth, though it can also cause inconveniences. What do you enjoy most? →0.4939
- (4) FSA: I think it is disappointing →-0.4939
- (5) NLP: Some may find rain disappointing, as outdoor plans can be disrupted. How do you cope? →-0.4939
- (6) FSA: Nope. Before I forget, do you like the rain? →0.4881
- (7) NLP: I appreciate the rain for its nourishing qualities. What about you, do you like it? →0.6369
- (8) FSA: I think it is incredible but terrible →-0.631
- (9) NLP: It's fascinating how the same weather can have such different perceptions. What makes it terrible? →0.1027
- (10) FSA: No. Before I forget, do you like the thunder? →-0.1531

5.1 Obtained Data

The FSA generates a text document for each strategy: mirroring and inverse mirroring. One iteration is defined as a pair consisting of an FSA bot response followed by an NLP bot reply. Each strategy is executed over 2,000 iterations (40 conversations of 50 iterations each), resulting in two separate documents containing a total of 4,000 responses each (8000 responses in total).

Once the data is collected, various graphs can be extracted to provide a clearer overview of the NLP bot's interaction with the FSA responses. Some examples of these visualizations are presented in Figures [2, 3]

5.1.1 Reading the Graphs. To aid in understanding the graph, some clarifications are provided below.

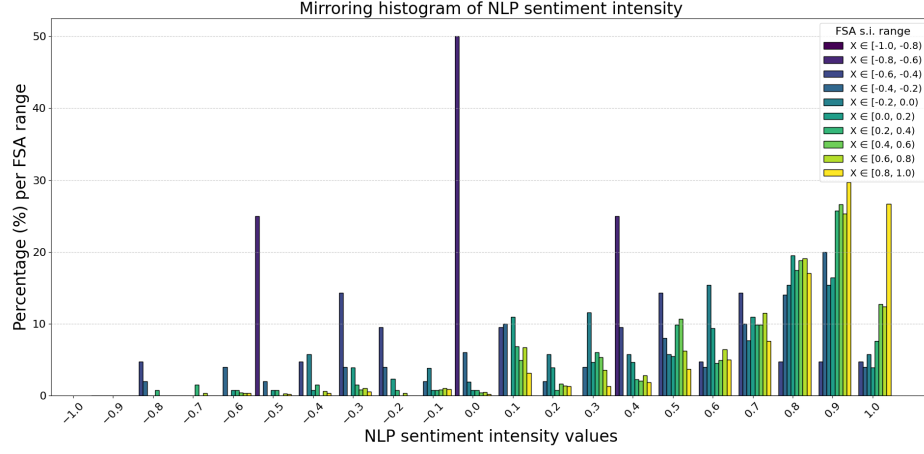


Figure 2: Mirroring strategy histogram by percentages

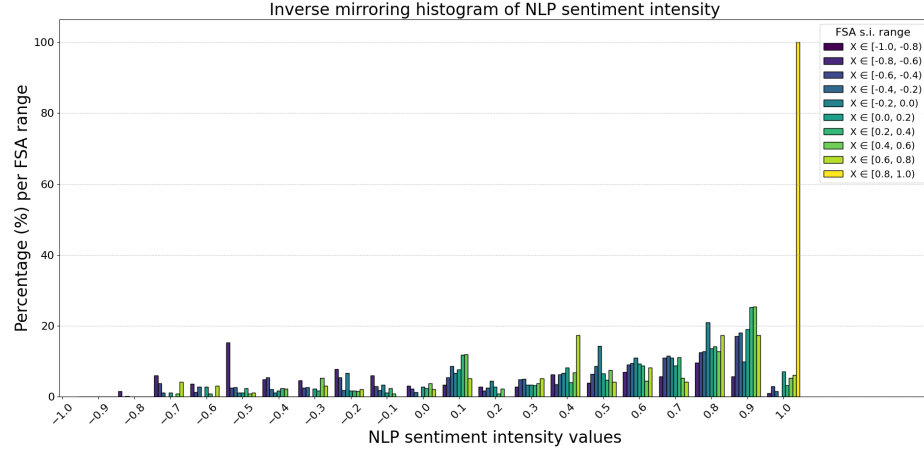


Figure 3: Inverse mirroring strategy histogram by percentages

Colors represent groups of sentiment intensity values derived from FSA responses, grouped into bins of width 0.1. Each FSA value is assigned to the next higher bin (e.g., 0.423 is grouped into the 0.5 bin). The lighter the color, the more positive the FSA response; conversely, darker colors represent more negative responses.

The position of each **bar** on the X axis should be interpreted as representing the closest rounded bin center rather than the exact sentiment intensity. Note that this does not mean that each individual sentiment value is rounded this way, the actual bin assignment still follows the approximation rule described for the FSA values (grouping into the next higher range).

This design choice is important because if each bar's position were based on the exact sentiment value, the graph would become unreadable, as many bars would end up overlapping each other.

5.2 Statistical Evaluation of the Results

To evaluate whether the NLP-bot's responses are influenced by the emotional intensity of the FSA's output, we conducted a statistical

analysis comparing sentiment intensity values from the mirroring and inverse mirroring experiments.

5.2.1 Mean and Standard Deviation. The mean sentiment intensity of the NLP-bot's responses under the mirroring strategy is **0.48** (with a standard deviation of **0.45**), while under the inverse mirroring strategy, it was **0.34** (standard deviation: **0.49**). Although these values alone do not allow for definitive conclusions, they provide useful descriptive statistics that inform further inferential analyses such as the T-test and Pearson correlation.

5.2.2 T-Test for Difference in Means. A two-sample independent T-test was conducted to determine whether the difference in mean sentiment intensities between the two strategies is statistically significant. The null hypothesis (H_0) assumes that the means are equal, whereas the alternative hypothesis (H_1) assumes a difference exists between them.

The test yielded a t-value of **10.98** and a corresponding p-value of 8.91×10^{-28} . Given a standard significance level of $\alpha = 0.05$, the

Algorithm 2: Generate Answer Closest to Target Sentiment Intensity for "What do you think?" related questions

Input: time, point, self.climate
Output: Sentence with sentiment closest to point

```

1: Function formIThinkAboutClimate(time, point):
2:   if climate  $\in$  simple_weather_terms or number of words in
      climate  $\neq 1$  then
3:     beginning  $\leftarrow$  "I think the " + climate + " is "
4:   else if climate == "it" then
5:     beginning  $\leftarrow$  "I think it is "
6:   else
7:     beginning  $\leftarrow$  "I think being " + climate + " is "
8:   most_accurate_try  $\leftarrow$  ("", -100)
9:   foreach word  $\in$  positive_adjectives  $\cup$  negative_adjectives
      do
10:    sentence  $\leftarrow$  beginning + word
11:    sentiment  $\leftarrow$  getSentiment(sentence)
12:    if  $|point - most\_accurate\_try[1]| > |point -$ 
      sentiment| then
13:      most_accurate_try  $\leftarrow$  (sentence, sentiment)
14:   foreach word1  $\in$  positive_adjectives do
15:     foreach word2  $\in$  positive_adjectives  $\cup$ 
      negative_adjectives do
16:       if word1  $\neq$  word2 then
17:         if word2  $\in$  negative_adjectives then
18:           sentence  $\leftarrow$  beginning + word1 + " but "
              + word2
19:         else
20:           sentence  $\leftarrow$  beginning + word1 + " and "
              + word2
21:         sentiment  $\leftarrow$  getSentiment(sentence)
22:         if  $|point - most\_accurate\_try[1]| > |point -$ 
          sentiment| then
23:           most_accurate_try  $\leftarrow$  (sentence,
              sentiment)
24:   return most_accurate_try[0]

```

p-value is far below the threshold ($8.91 \times 10^{-28} < 0.05$), allowing us to confidently reject the null hypothesis. This result indicates that the difference in sentiment responses under the two strategies is statistically significant.

5.2.3 Pearson Correlation. To further investigate the relationship between the FSA's emotional intensity and the NLP-bot's sentiment output, we calculated the Pearson correlation coefficient. This test assesses the strength and direction of a linear relationship between two continuous variables.

The analysis yielded a correlation coefficient of **0.38** with a p-value of 3.42×10^{-136} . This suggests a moderate positive correlation between the sentiment intensities of the FSA and the NLP-bot. In other words, the more positive (or negative) the FSA's sentiment, the more positive (or negative) the NLP-bot's reply tends to be. As

the p-value is significantly lower than the conventional threshold of 0.05, we reject the null hypothesis of no correlation and conclude that the observed relationship is statistically significant.

6 DISCUSSION

6.1 Limitations

It is important to note that this study focused exclusively on interactions with a single NLP model, ChatGPT_3.5_turbo, within a constrained conversational weather-related dialogue domain. While this allows for greater control and replicability of experiments, it also limits the generalization of the findings. Results may differ when applied to other chatbot architectures, such as rule-based systems or transformer models with different training objectives, as well as in conversations involving more emotionally complex or sensitive topics. Future work should extend this approach to a broader set of models and domains in order to assess the robustness of the observed patterns.

6.2 Threads to validity

Any variant of a question that uses contractions or abbreviations can also be correctly identified by the FSA. For example, it is capable of recognizing both "How is the weather?" and its contracted form "How's the weather?", even though the initial prompt encourages avoiding contractions.

The main reason why an error rate for the FSA's emotional intensity responses was not included is due to the inherent limitations of sentiment approximation in short, generic responses. In many cases, accurately matching the target intensity value is simply not feasible without resorting to overly long or complex sentences.

This trade off led to the decision not to define specific error cases: either generating a perfectly matched response would be computationally expensive, or deliberately forcing imperfect matches would clutter the results with noisy, artificial errors. In any case, this limitation does not compromise the core purpose of the experiment, as the primary focus is on observing how the NLP bot interprets and responds to emotional cues, rather than how precisely the FSA can mirror a specific emotional target.

6.3 Analysis of results

The results presented in this study suggest that the NLP model (ChatGPT) is not only capable of producing emotionally coherent responses but also shows sensitivity to the emotional intensity of the inputs it receives. Visual evidence, as shown in Fig. 2 and 3, indicates a general tendency of the model to respond with positive sentiment, regardless of the input strategy. However, when analyzed in more detail, a pattern emerges: the lighter the color (representing higher FSA sentiment), the more concentrated the bars are toward the positive side of the sentiment spectrum. This suggests that the sentiment of the FSA influences the sentiment expressed in the NLP-bot's response.

To substantiate these observations, a statistical evaluation was conducted. Descriptive statistics showed a noticeable difference in the average sentiment intensity of the NLP-bot under each strategy: **0.48** (SD = **0.45**) for mirroring and **0.34** (SD = **0.49**) for inverse mirroring. While these differences suggest a trend, inferential statistics provide stronger evidence.

A two-sample independent T-test was performed to evaluate whether the observed difference in means was statistically significant. The test yielded a t-value of **10.98** with a p-value of 8.91×10^{-28} , leading to the rejection of the null hypothesis at a 5% significance level. This confirms that the difference in sentiment between the two strategies is not due to random variation and is statistically significant.

Additionally, the Pearson correlation coefficient was calculated to examine the linear relationship between the sentiment intensity of the FSA's responses and those of the NLP-bot. The coefficient obtained was **0.38** with a p-value of 3.42×10^{-136} . This reflects a moderate positive correlation, suggesting that the sentiment expressed by the NLP-bot tends to increase as the sentiment of the input increases. Once again, the extremely low p-value supports the statistical significance of this correlation.

Taken together, the graphical analysis and the statistical evaluation provide strong evidence that the NLP-bot is sensitive to the emotional intensity of its conversational partner. While it exhibits a general bias toward positive sentiment, the intensity and valence of the input do exert a measurable influence on its output. This behavior may be interpreted as a form of emotional alignment or "positive mirroring," wherein the model attempts to match the affective tone of the input while maintaining a preference for positive emotional expression.

These findings contribute to a growing understanding of how large language models handle emotional context and offer valuable insights for the development of emotionally intelligent conversational agents.

This suggests that ChatGPT is not only capable of generating emotionally coherent answers, but also of adapting its sentiment in response to the emotional context of the conversation.

7 CONCLUSION

7.1 First research question

What is the capacity of NLP bots to detect and respond to emotional intensity in user inputs? The results indicate that NLP bots such as ChatGPT possess a limited but measurable capacity to detect and respond to emotional intensity. While the model tends to favor positive responses overall, it shows a consistent pattern of adjusting its sentiment based on the valence and strength of the input it receives. This was evidenced by the moderate positive correlation (0.38) between the sentiment intensity of the FSA responses and those of the bot, indicating that ChatGPT can recognize emotional cues and reflect them in its replies.

7.2 Second research question

In what ways do NLP bots adjust their strategies (such as mirroring) based on the emotional strength of the messages they receive? ChatGPT demonstrated the ability to adapt its output in alignment with the emotional strategy of the conversation. Under the mirroring condition, the bot responded with sentiment values closer to those of the FSA inputs, while in the inverse mirroring condition, the sentiment shifted accordingly in the opposite direction. These results show that the bot does not rely on a fixed strategy but rather dynamically adjusts its tone depending on the affective input it receives, suggesting an emergent emotional alignment mechanism.

7.3 Third research question

To what extent do NLP bots exhibit a tendency to generate more positive or negative responses when aligned with emotional input from users? While the NLP bot adapts its responses based on emotional cues, it consistently leans toward more positive expressions. Even when receiving negative input, ChatGPT often softened its tone. Nevertheless, the bot still reflects input sentiment to a significant degree, indicating a nuanced balance between alignment and positivity.

7.4 Final conclusion

In summary, this research confirms that NLP bots like ChatGPT are not only capable of detecting emotional intensity but also of adjusting their responses in ways that reflect the affective tone of the conversation. The use of a Finite State Automaton allowed for a controlled and scalable evaluation, and the results supported by both visual analysis and statistical tests demonstrate a significant relationship between input sentiment and bot output. While the bot exhibits a general bias toward positivity, it nonetheless mirrors emotional intensity to a meaningful extent. These findings provide valuable insights for future improvements in emotionally intelligent conversational agents and highlight the growing potential of NLP systems in socially and emotionally sensitive applications.

8 FURTHER WORK

To continue advancing research in the field of NLP bots and their sensitivity to emotional intensity, several directions for future work can be proposed. First, the capabilities of the Finite State Automaton (FSA) could be extended to handle conversations on a broader range of topics beyond weather. This would allow for the evaluation of NLP bot responsiveness in more diverse and emotionally rich contexts.

Additionally, testing multiple NLP models such as other versions of ChatGPT or different conversational agents would enable a comparative analysis of their emotional alignment abilities. This would help determine whether the observed behaviors are model-specific or generalizable across platforms.

Finally, a more refined analysis could involve classifying each interaction based on the predominant emotion expressed in the FSA response (e.g. joy, sadness or anger), and then evaluating how the bot responds to each specific emotional category. This approach would provide deeper insights into how NLP bots handle not just sentiment intensity in general, but emotional nuances across different affective states.

REFERENCES

- [1] Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications* 2 (2020), 100006. <https://doi.org/10.1016/j.mlwa.2020.100006>
- [2] Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Álvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3971–3984. <https://aclanthology.org/2020.emnlp-main.326>
- [3] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. EMMA: An Emotion-Aware Wellbeing Chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7. <https://doi.org/10.1109/ACII.2019.8925455>
- [4] C.J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8. 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- [5] Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2017. Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 1–39. <https://doi.org/10.1145/3124420>
- [6] Tingting Liu, Salvatore Giorgi, Ankit Aich, Allison Lahnala, Brenda Curtis, Lyle Ungar, and João Sedoc. 2025. The Illusion of Empathy: How AI Chatbots Shape Conversation Perception. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 13 (Apr. 2025), 14327–14335. <https://doi.org/10.1609/aaai.v39i13.33569>
- [7] Diana Maynard and Kalina Bontcheva. 2014. Challenges of Evaluating Sentiment Analysis Tools on Social Media. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. <https://eprints.whiterose.ac.uk/id/eprint/98769/>
- [8] Kristina Schaaff, Caroline Reinig, and Tim Schlippe. 2023. Exploring ChatGPT’s Empathic Abilities. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. <https://doi.org/10.1109/ACII59096.2023.10388208>
- [9] Mark Utting, Alexander Pretschner, and Bruno Legeard. 2012. A Taxonomy of Model-Based Testing Approaches. *Software Testing, Verification and Reliability* 22, 5 (2012), 297–312. <https://doi.org/10.1002/stvr.456>
- [10] Douglas Youvan. 2024. Understanding Sentiment Analysis with VADER: A Comprehensive Overview and Application. <https://doi.org/10.13140/RG.2.2.33567.98726>