Shelley: A Crowd-sourced Collaborative Horror Writer

Pinar Yanardag Delul*
yanardag.pinar@gmail.com
webmaster@marysville-ohio.com
Bogazici University
Dublin, Ohio, Turkey

Manuel Cebrian*
cebrian@mpib-berlin.mpg.de
webmaster@marysville-ohio.com
Center for Humans and Machines, MPI
Dublin, Ohio, Germany

Iyad Rahwan*
rahwan@mpib-berlin.mpg.de
webmaster@marysville-ohio.com
Center for Humans and Machines, MPI
Dublin, Ohio, Germany

ABSTRACT

Fear induction in the form of stories and visual images pervades the history of human culture. Creating a visceral emotion such as fear remains one of the cornerstones of human creativity. As artificial intelligence makes strides in solving challenging analytical problems like chess and Go, an important question still remains: can machines induce extreme human emotions, such as *fear*? In this work, we propose a deep-learning based collaborative horror writer that collaboratively writes scary stories with people on Twitter. We deploy our system as a bot on Twitter that regularly generates and posts new stories on Twitter, and invites users to participate. Users who interact with the stories produce multiple storylines originating from the same tweet, thereby creating a tree-based story structure. We further perform a validation study on $n = 105$ subjects to verify whether the generated stories psychologically move people on psychometrically validated measures of effect and anxiety such as I-PANAS-SF [43] and STAI-SF [26]. Our experiments show that 1) stories generated by our bot as well as the stories generated collaboratively between our bot and Twitter users produced statistically significant increases in negative affect and state anxiety compared to the control condition, and 2) collaborated stories are more successful in terms of increasing negative affect and state anxiety than the machine-generated ones. Furthermore, we make three novel datasets used in our framework publicly available at https://github.com/catlab-team/shelley for encouraging further research on this topic.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Psychology**.

KEYWORDS

crow-sourcing, human-AI collaboration, deep learning, emotion, fear, datasets

C&C '21, June 22–23, 2021, Virtual Event, Italy

ACM ISBN 978-1-4503-8376-9/21/06…$15.00
https://doi.org/10.1145/3450741.3465251

ACM Reference Format:
Pinar Yanardag Delul, Manuel Cebrian, and Iyad Rahwan. 2021. Shelley: A Crowd-sourced Collaborative Horror Writer. In *Creativity and Cognition (C&C '21), June 22–23, 2021, Virtual Event, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3450741.3465251

1 INTRODUCTION

The eruption of Mount Tambora in 1815 caused utterly strange meteorological phenomena during the subsequent spring and summer of 1816. The world experienced a seemingly never-ending winter, popularly known as the "Year Without a Summer" [53]. Among the many cultural consequences of this episode, a remarkable one involved a group of notable writers retreating at Villa Diodati, a Mansion at Lake Geneva. Being house-bound due to the long winter, some of the best writers of the Victorian era, Mary Shelley, John William Polidori, Lord Byron, among others, held an informal competition to see who could produce the scariest story ever written [50]. Shelley created the iconic figure of Dr. Frankenstein [51]; Polidori planted the seed of Vampirism [52]; and Byron, in his poem Darkness [49], narrated by the last man on earth – produced the foundational piece of the apocalyptic horror genre. Though excellent at their craft, these writers were not unusual in their desire to devise ways to terrify their fellow humans. Such attempts at fear induction – taking the form of stories and visual images – pervade the history of human culture. Creating a visceral emotion such as fear remains one of the cornerstones of human creativity. As Artificial Intelligence (AI) makes strides in solving challenging analytical problems like checkers [37], chess [39] or video-games [47] and defeating the world's best Go and chess players [38, 39] society takes solace in the implicit belief that the subset of human tasks that rely on the understanding, managing, and inducing human emotions are safe from machine overtake. But are they? Can computers learn to create scary stories and collaborate with human authors to create even scarier ones?

Automated story generation is a popular research topic in the natural language generation community. With the recent advancements in deep learning, researchers focused on the intersection between natural language processing and human-computer interaction for creative story generation. Several tools have recently been proposed to directly collaborate with human authors in order to provide automated support for story writing [12, 14, 17, 35]. These approaches range from providing cue phrases for usersduring the generation process [7] to suggesting story continuations for users [6]. Supported by the quality of the generated text, interactive story generation tools have already been used by professional authors. For instance, novelists use generative language models to either finish their sentences or to generate the next paragraph of text in

order to generate ideas for scenes and characters or antagonistic suggestions to improve their writing [10].

In this paper, we explore collaborative horror story writing using a neural network trained on a large-scale short-horror story corpus collected from subreddit */r/nosleep*[1]. What is unique about our approach is that we explore Twitter as a medium for collaborative story writing to encourage large-scale participation and our approach leads to several alternative stories originating from a single story due to the multi-threading structure of Twitter conversations. Our story-writer bot:

(1) generates and posts story snippets on Twitter in regular intervals and invites users to continue its stories,
(2) responds replies from users by generating story continuations conditioned on user-provided context,
(3) responds to new stories posted by the users in order to support user-initiated stories.

Over a period of two weeks, our Twitter bot gained the attention of over 6K Twitter followers, and generated over 500 human-AI collaborated stories. An example story with two alternative storylines can be seen above (STORY I). Text tagged with [START] is a story generated by our model (shown in *italic* font), and [THREAD I], [THREAD II] and [THREAD III] are three alternative stories continued by different Twitter users[2] (shown in ***bold italic***. After users continue the initial [START] tweet by sending a reply, our bot generates a continuation conditioned on the context created so far (e.g. initial story snippet + the continuation posted by the user). Our contributions are as follows:

- We propose the task of horror story generation, and introduce a new way for collaborative story writing using Twitter as a medium,
- We run a validation study I-PANAS-SF and STAI-SF metrics and show that the generated stories as well as collaborated stories produced statistically significant increases in negative affect and state anxiety compared to the control text.
- We share three novel datasets to encourage further research on this topic: 1) a large-scale dataset of 134K horror stories collected from the Reddit platform 2) a dataset of 300 generated stories labeled by Amazon Mechanical Turk participants on a Likert scale indicating the scariness of the text 3) a tree-based dataset produced by our bot and Twitter users that includes multiple threads originating from initial stories.

The rest of the paper is organized as follows. Section 2 discusses related work in story generation and crowd-sourced AI tools. Section 3 discusses the technical details of the Twitter bot. Section 4 discusses validation experiments we performed. Section 5 concludes the paper.

2 RELATED WORK

In this section, we discuss related work in story generation and collaborative tools in deep learning.

[1]Nosleep subreddit: http://www.reddit.com/r/nosleep.
[2]See the relevant thread: https://twitter.com/shelley_ai/status/923308554852995074.

2.1 Story generation

Story generation is a popular problem in natural language generation with efforts as early as the 1970s [20, 27] where it was viewed as a symbolic planning task. Later approaches adopted case-based reasoning (CBR) [13], or domain-model based approaches [22] by crowd-sourcing a corpus of narrative examples and generating stories by sampling from a domain model. Systems such as MakeBelieve [23] use commonsense rules for action sequences from a knowledge database.

Recurrent neural network (RNN) based models are employed to generate stories based on the next character, word or sentence. [36] uses an LSTM-based network to generate stories, [30] proposes an RNN-based approach called Story Scrambler that generates new stories based on inputted stories. [15] generates stories from sequences of short narrations by using a sequence-to-sequence RNN architecture. Recent work on story generation often uses sequence-to-sequence [41] or attention [25] based models. [12] proposed a system that builds coherent and fluent passages of text based on a premise and uses a hierarchical approach based on a fusion mechanism [40].

Early examples of interactive story generation can be found from Choose Your Own Adventure (CYOA) novels [28] where users can control the story narrative by choosing specific pages using a branching story graph. Interactive story generation using machine learning became a popular research area [34] in which users influence storylines through their actions. [7] focuses on the task of interactive story generation, where the user provides the model mid-level sentence abstractions in the form of cue phrases during the generation process. [6] proposed a system called STORIUM where human authors query a model for suggested story continuations and edit them.

One of the early works that provide automated support for story writing is Say Anything system [42] where users and computer take turns in writing sentences of a fictional narrative via sentences from a collection of a large-scale story dataset. [18] explored a collaborative writing system called Ensemble that brings together a group of people to write collaboratively in order to create a single story. Each story had a lead author and contributors submitted alternate versions of a scene which is then rated and the winning scene is chosen by the lead author. [35] proposed an application that generates suggestions for the next sentence in a story where users can modify or delete suggestions based on their choice. [17] proposes a text prediction system trained on a specific theme in order to explore data-driven creativity and productivity. [12] explores coherent text generation using a hierarchical model based on human-written stories paired with writing prompts from an online forum. [14] proposes a model for generating story endings for a given story context while handling implicit information to keep the story coherent. Selecting diverse prompts from generated outputs is also a related research problem. [11] introduce an automatic prompt selection approach using a language model embedding to direct users towards diverse prompts to maximize diversity.

To the best of our knowledge, our work is among the first approach that uses Twitter as a crowd-sourcing platform for collaborative story writing with a deep learning based system.

STORY I – Three alternative storylines originated from the same tweet (italic text: AI, bold text: human).

[START] *When I heard the phone ring again, I ran to the stairs. As I was running down the stairs, I started to hear crying. I shone my phone around the corner of the staircase and saw the crying baby getting closer. I crawled over to it and kicked it as hard as I could. The crying from the stairs turned into a soft metallic sound.*

→ [THREAD I] ***I turned back towards the hallway I came from, nothing seemed to be the same, I felt lost, things had been moved from their place*** *and the only thing that caught my attention was the light from the hallway. At this point I could not believe what I was seeing.*

→ [THREAD II] ***Something was missing. I felt incomplete but couldn't really grasp what was wrong.*** *I looked at the door that led to the basement and saw a pair of eyes staring back at me.*

→ [THREAD III] ***All of a sudden, there was no sound. There was pin drop silence. Then, the shrill cry of the baby rattled my bones.*** *I could feel my shoulder shaking, and I felt like I was being stabbed in the neck and everything went black.*

2.2 Creative AI Tools

The field of machine learning has recently gained an immense amount of attention due to breakthrough results in several important tasks. This success also encouraged researchers to focus on generative models with crowd-sourcing efforts for creative applications of deep learning.

Computer vision based platforms such as Deep Dream Generator [3] enabled users to experiment with deep learning algorithms for creativity. Neural style transfer algorithm [21] allowed users to transfer painting styles to a given image [4]. Other creative AI tools include music-based platforms such as Magenta [5] which offers a large collection of music-based tools such as a recurrent neural network (RNN) based system that generates notes based on the drum beat or melody provided by the users.

Text-based platforms such as Botnik [2] and GPT-2 [32] also heavily explored for creative and collaborative writing. Botnik offers a *keyboard*-based interface where users can collaboratively create AI-assisted text-based content. GPT-2 [33] and GPT-3 models [9] benefit from large-scale datasets and enabled users to create a large variety of creative work ranging from novels [1] to poetry [8].

Computationally creative Twitter bots in the wild are also explored in several studies [44] such as [45] which explores Twitter as a medium for automated wit via a Twitter bot named *@MetaphorMagnet*. [29] proposes a bot that tweets poems inspired by Twitter trends. It paraphrases text by Twitter users or produces new text fragments by extracting or inferring semantic relations. [19] builds poems from tweets scored according to a specific criteria such as reaction, meaning or craft. [46] proposes a bot that generates riddles about celebrities by retrieving content related to the celebrity and generating analogies based on the relevant attributes. The generated riddles are then shared with users and their answers were evaluated by the bot.

3 METHODOLOGY

We deployed our model as a bot on Twitter at http://twitter.com/shelley_ai a week prior to Halloween on October 2017 and organically grew to an audience of over $6K$ followers. Over the period of 10 days, users collaboratively wrote over 500 stories on Twitter. Our bot is responsible for three primary tasks. First, it generates and posts a new story every hour, and invites its followers to participate. Second, it tracks the responses sent by the users and automatically generates and posts new continuations. Third, it responds to mentions (addressed to **@shelley_ai** on Twitter) for new stories initiated by Twitter users and responds to them with story continuations it generates using the context created so far. Figure 1 illustrates the main components of our framework. *Story Generator* component is responsible for generating stories or continuation for stories with multiple options. *Story Ranker* component ranks the options based on *scariness*, *Story Poster* posted the top option on Twitter, and *Response Collector* collects the user replies and sends the context gathered so far back to Story Generator. We discuss each component below.

STORY GENERATOR. Our Story Generator component consists of a recurrent neural network based architecture trained on a large-scale short horror stories collection we crawled from Reddit's **r/nosleep** subreddit. NoSleep is a community of $14.6M$ users[3] where users post short stories they write. Each post must be original and must be a horror story in order to be featured on the subreddit. The stories have varying lengths and have a large diversity in terms of topic. We scraped 7 years of posts between 2010-2017 using Reddit API and cleaned the dataset by removing deleted posts, automated posts or announcements. The final dataset consists of 134,500 stories, having 133M words and 686M text characters. The most common keywords that occur in the dataset are shown in Figure 2. The Story Generator model is designed as a two-layer character-based recurrent neural network with 1024 hidden units and a sequence length of 128. We use dropout as 0.25 and batch size as 512 and trained the model for 150K iterations. The model is responsible for generating two different types of text: *story starters* that are used for initiating stories on Twitter, and *story continuations* for continuing a given story content created so far. For generating *story starters*, the first character is chosen randomly, and next 256 characters are generated based on the trained model. The other type of generation is *conditional* where we take the initial story as well as the replies posted by the users (if any) and start off the generation process with this context.

For both cases, we randomly generate 10 different story options using our model and send the candidate batch to *Story Ranker* in order to select the best option (see Section 3). While generating the candidates, we apply the softmax temperature trick [16]. Temperature t is a hyperparameter often used in recurrent neural networks to control the randomness of the generation. This parameter is used for scaling the logits before applying softmax where larger values

[3]Last accessed on Feb 1, 2021.

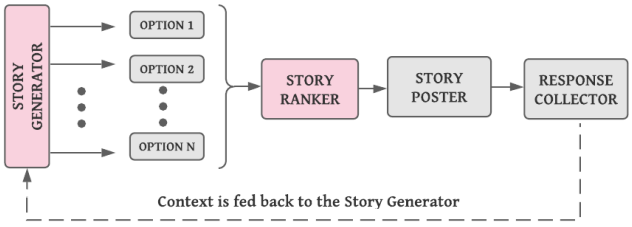

Figure 1: An illustration of our framework. Story Generator generates multiple story options, and Story Ranker scores each option based on their scariness. Story Poster posts the top choice on Twitter, and Response Collector tracks the replies from the users and sends them back to the Story Generator in order to generate a continuation for the story conditioned on the context gathered so far.

Figure 2: Top words from the r/nosleep dataset are shown. Size of the words are correlated with their frequency.

produce a softer probability distribution over classes and results in more diversity. We vary the temperature parameter between $t = \{0.4, \ldots, 0.9\}$ to diversify the outputs. **STORY II** box shows an example generation. **START** is a story starter generated by our model, and **OPTION 1-3** are candidate continuations for the text with different temperature parameters. Note that low temperatures tend to generate repeated content while keeping the context close to the original text (e.g. **TEMP=0.4**), while higher temperatures generate more diverse content but making the context far from the original (e.g. **TEMP=0.7**).

STORY RANKER. *Story Ranker* component takes 10 candidate continuations generated by the *Story Generator* as input and outputs the *scariest* one to be sent on Twitter. In order to assess how *scary* a given text is, we train an RNN-based classifier [24] on a labeled dataset obtained via Amazon Mturk [4]. Participants were shown random chunks of text from our model with a temperature parameter varying between 0.5 and 0.9, and asked to rate the texts on a scale of 1 (not-scary) to 5 (very scary). In total, 1,739 snippets were voted by $n = 300$ users which are used to train the classifier model.

STORY POSTER. Our bot posted a new story starter on Twitter once per hour over a period of 10 days. Each new post consists of up to 3 tweets and ended with a special *#yourturn* hashtag that prompts users to continue the story. Story starters are picked among 10 candidates by the *Story Ranker* based on their *scariness* score.

RESPONSE COLLECTOR. Twitter users collaborate and continue the stories by sending one or more replies under a particular tweet and include *#end* hashtag to indicate the end of their story. Once the bot encounters the hashtag, it generates a response and posts it back. For instance, the following is a story example where the user continued the story from the first tweet ([START] and [START (cont)] are two tweets posted by our bot and [THREAD] is created by a user who continued the story from the first tweet). Due to the length limit, the bot sends stories in 1-3 tweets and users can choose to continue from a particular part of the story.

The inherent threading property on Twitter results in a tree structure where a single tweet can result in multiple storylines. Figure 3 shows an example of 7 storylines consist of 36 tweets (Tweets $T_1, \ldots, T_{36}$). Some responses consist of multiple tweets such as $T_{23,24,25}$. While most of the storylines continued in parallel originating from the initial tweet, some storylines branched into two after a certain point (such as $T_{26,27}$). **STORY IV** shows the corresponding stories in the tree structure.

4 VALIDATION STUDY

We run a validation study in order to understand whether the generated stories psychologically scare people. We ran a controlled experiment on Amazon's Mechanical Turk on $n = 105$ subjects by employing psychometrically validated measures of effect and anxiety. We randomly assigned the subjects to three treatment groups: **Control, Bot, Bot+Human**. The *Control* group consisted of ten stories that are randomly selected from Gutenberg Short Stories Collection [31]. The purpose of the control group is to represent neutral stories. For *Bot* and *Bot+Human* conditions, we first generated 10 story starters using our deep learning model. Then, we followed two different approaches to generate stories: the first approach takes story starters and automatically generates continuations with our model (*Bot* condition). The second approach takes

[4]Participants were required to be located in the USA and required to have an approval rate higher than 95%. Participants were paid 0.50 USD for completing the study.

STORY II – Multiple story options where t represents the temperature parameter.

[START] *The air was almost dead... it sounded like a gentle-silent demonic tone I had never seen before. I wanted her to be one of my children.*

→ **[OPTION 1][t=0.4]** *I stood there in terror for a moment, then started to scream. I was so scared that I didn't know what to do. I had to know that something was wrong. I wanted to scream and run but I couldn't. I couldn't scream, I couldn't move. I couldn't move. I couldn't move.*

→ **[OPTION 2][t=0.5]** *I didn't know what to do. I just wanted to get out of here. I wanted to make sure all of this was a good idea. The children were laughing too. They couldn't even close their eyes to make the words feel like any other soul would. Then they started to cry.*

→ **[OPTION 3][t=0.7]** *I felt so relieved to find out that the next day I was going to try to keep the shadows on back up. The more I tried, the more it started to turn around and get the better of me.*

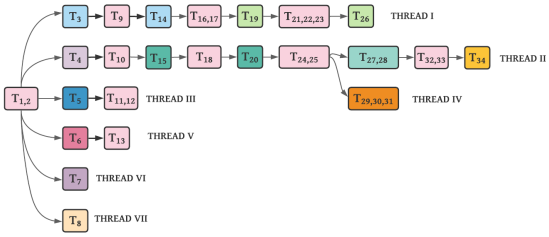

Figure 3: An example tree of 7 stories (see STORY III box) originated from a single story is shown. T_i represents the tweet ID, and each row in the figure shows an alternative thread resulting from $T_{1,2}$ root tweet. Boxes in light pink are tweets posted by our bot while other colors represent Twitter users (the same color code is used for the same users).

STORY III –Users can choose to continue the stories from any tweet (italic text: AI, bold text: human).

[START] *I heard the shift in his voice and felt his breath on my neck. I ran and screamed and looked back hoping to wake my parents when I saw*

→ **[START (cont)]** *him standing there. He looked at me and shook his head.*

→ **[THREAD] his eyes, eyes I'd spent countless hours gazing into, dreaming, were empty sockets. His mouth, torn in anguish, uttered a moan.**

the same story starter as the *Bot* condition and continues stories collaboratively with Twitter users. All three conditions have the same length (with ±10 characters to allow word completion) in order to avoid biases based on length. STORY V box shows examples of stories for each control group[5]. [CONDITION I] and [CONDITION II] are created with initial story starter [STARTER] while story from [CONDITION III] is directly taken from an existing Gutenberg story[6]. Each participant is shown one of the three conditions randomly where they are asked to read the corresponding short story and then asked to answer a questionnaire to understand how they were affected by the content they read. We used two measures that are commonly used in psychology: I-PANAS-SF and STAI-SF. I-PANAS-SF is the Positive and Negative Affect Schedule and derived from the original twenty PANAS study [48] and measures dimensions of positive and negative affect.

I-PANAS-SF consists of ten items including five positive affective states: *Active, Determined, Attentive, Inspired*, and *Alert* and five negative affective states: *Afraid, Nervous, Upset, Hostile*, and *Ashamed*. Participants are asked to respond to the positive and negative states after reading the particular story shown based on the condition they were assigned to. They are allowed to answer based on a 5-point scale ranging from *Very slightly or not at all* (1 point), *A little* (2 points), *Moderately* (3 points), *Quite a bit* (4 points), *Extremely* (5 points). The *Positive Effect* is calculated as the total score the participant gave for *Active, Determined, Attentive, Inspired*, and *Alert* items, while *Negative Effect* is calculated as the total score that participants gave for *Afraid, Nervous, Upset, Hostile*, and *Ashamed* items. The second metric is a shortened version of the State-Trait Anxiety Inventory (STAI-SF) which measures the state anxiety of the participants. Participants can respond to six items assessing the degree that patients feel *Calm, Tense, Upset, Relaxed, Content*, and *Worried*. It uses a 4-point Likert scale ranging from *Not at all* (1 point), *Somewhat* (2 points), *Moderately* (3 points), and *Very much* (4 points). The scores for all items are summed up where higher scores are positively correlated with higher anxiety. For both I-PANAS-SF and STAI-SF, the order of the outcome measures is randomized in order to avoid any ordering effects.

Negative Effect, Positive Effect, and State Anxiety results can be seen from Figure 4. The results of our experiment indicate that both AI-generated (*Bot* condition) and human-AI collaborated stories (*Bot+Human* condition) produced significant increases in negative affect and state anxiety measures. Participants in *Bot* and *Bot+Human* conditions chose significantly higher scores on State Anxiety measure (STAI-SF) and Negative Affect Schedule (I-PANAS-SF-Neg) comparing to the control group. We run a paired T-test between scores of *Bot+Human* and *Control* conditions for STAI-SF scores and find out a $t-$statistic of 4.306 with a p-value of $6.49e-05$ which suggests that the human-AI collaborated stories were able to significantly increase anxiety and negative affect. We also compared

[5]Only short snippets are shown here due to the length.
[6]HAMLIN GARLAND, A Camping Trip: http://www.gutenberg.org/files/20831/20831-h/20831-h.htm

STORY IV –Storylines originated from a single story (see Figure 3).

[START] *I then looked out my window to the side of my house and I could see the figure of a man standing in the woods. I froze and then I ran 1after him. I ran and ran and ran and ran, and then I realized he was*

→ [THREAD I] ***not a man at all. But something far worse. It stopped to look back at me and the breath caught in my throat.*** *It was a tall black humanoid figure with no facial features and i didn't know why it was staring.* ***I realised too late it was probing me, trying to find a way into my mind. An awful alien voice filled my head and I couldn't run*** from it anymore. It felt like forever, and I felt something tickle me, a burning ***that I knew would never stop. I understood now its sole purpose: not just killing me but erasing all memory of my existence*** . I was unable to move and I stared at this thing for a moment and then I felt its skin was cold and pale and then it felt like a shifting body in my lungs was trying to stay in my soul.

→ [THREAD II] ***He was slowly taking over my body.. I could feel his alien presence inside me. He chosen me to be his host; his human companion 10 while he is going to take over the world.*** The sound of my last breath burned to a crack in his body. I could feel it raise above my body. I was terrified by what was 12coming. I looked up at my "body" and saw myself floating ***In my last conscious thought, I felt sorry for the world, for it will be over soon, and felt proud of the alien race I belonged to.***

→ [THREAD III] ***The alien buzzing in my head, the coldness in my lungs, I struggled to maintain my own senses. All the while seeing colors and hearing 10 tones I couldn't recognize. Combined with the buzzing in my mind it was too much. As if these revelations were supposed to be payment 11 for the deletion of my existence. The faceless thing beckoned me closer, and despite my agony, I could not refuse.***

→ [THREAD IV] ***going deeper into the woods, and I followed him without noticing that the path had ended. It was pitch black and I couldn't see...*** *my body started to feel like I was going to cut myself from the inside out. I couldn't see him anymore.*

→ [THREAD V] ***He was my dad but he does not have eyes!*** *No eyes or sockets or anything, but I heard a voice, deep and warm. He was crying as if he was being shaken from the inside out*

→ [THREAD VI] ***someone I knew but he was wearing a different face. His eyes were cold but they knew me. When he looked at me what did he see?*** *I just recognized him! it was him! He had a grin on his face*

→ [THREAD VII] ***Wasn't getting closer but rather farther away from me. Distressed, I spun around to something more horrifying. Nothingness.***

→ [THREAD VII] ***An older me.***

STORY V –Example stories for the three conditions used in our validation study.

[START] *it smiles. it is smiling now. i cannot see its mouth. it has no eyes at all. it has no mouth, and no nose. it is smiling.*

→ **[CONDITION I: BOT+HUMAN]** it smiles because it knows i am here. its smooth face slowly turns toward me, looking at me without any eyes. i cant see its face, but it is holding a dead rotting corpse in its hand. as it lifts the corpse towards me, i stumble backwards into the corner. it stops suddenly, sensing that i am vulnerable. i look back at the corpse, and there is blood everywhere. i start to panic. i try to hide from the creature but i cant reach it. (...)

→ **[CONDITION II: BOT]** its mouth is moving and bleeding, and its mouth is darkening, with an orange crooked smile. it is a smirkish smile caught on this things lips. and at both the thing it has, It is a puppet. even from just behind it, does not have a face, and more like a cat. it is back, and staring at me. it continues sprinting taking the front of the stairs. breathing in protest as the hissing becomes louder. the sounds have become more close together (...)

[CONDITION III: CONTROL] It was the fifteenth of June, and the sun glazed down upon the dry cornfield as if it had a spite against Lincoln Stewart, who was riding a gayly painted new sulky corn-plow, guiding the shovels with his feet. The corn was about knee-high and rustled softly, almost as if whispering, not yet large enough to speak aloud. Working all day in a level field like this, with the sun burning one's neck brown as a leather glove, is apt to make one dream of cool river pools, where the water snakes wiggle to and fro, and the kingfishers fly above the bright ripples (...)

the *Bot* and *Control* conditions and found a $t-$statistic of 3.197 with a $p-$value of 0.002, which suggests that stories that are completely generated by the *Bot* also increased anxiety and negative affect in a statistically significant way.

We observe a similar trend for I-PANAS-SF (Neg) metric, where we obtained a $t-$statistic of 3.85 and a $p-$value of 0.0003 between *Bot+Human* and *Control* groups, and a $t-$statistic of 3.07 and a $p-$value of 0.003 for *Bot* and *Control* groups.

Moreover, we investigated the differences between conditions for Positive Affect (see Figure 5), and find out a $t-$statistic of -0.24 ($p = 0.80$) between *Bot+Human* and *Control* groups, and a $t-$statistic of -0.57 ($p = 0.56$) between *Bot* and *Control* groups. Therefore, even though generated stories were also to reduce positive affect compared to the control group, they did not significantly differ.

5 CONCLUSION

In this work, we explored the potential of inducing fear by generating *scary* stories. We launch our framework as a Twitter bot in order to enable large-scale participation which resulted in over 500 human-AI collaborated horror stories. Our bot generates new stories and prompts users to continue them, as well as generating continuations for stories initiated or participated by Twitter users. We run a validation study where we performed a controlled experiment on $n = 105$ subjects on Amazon's Mechanical Turk where we verify whether the generated stories psychologically move people on psychometrically validated measures of effect and anxiety such as I-PANAS-SF and STAI-SF. Our exploratory results and validation experiment suggests that deep learning and generative algorithms have a significant potential for inducing emotions. Furthermore, we

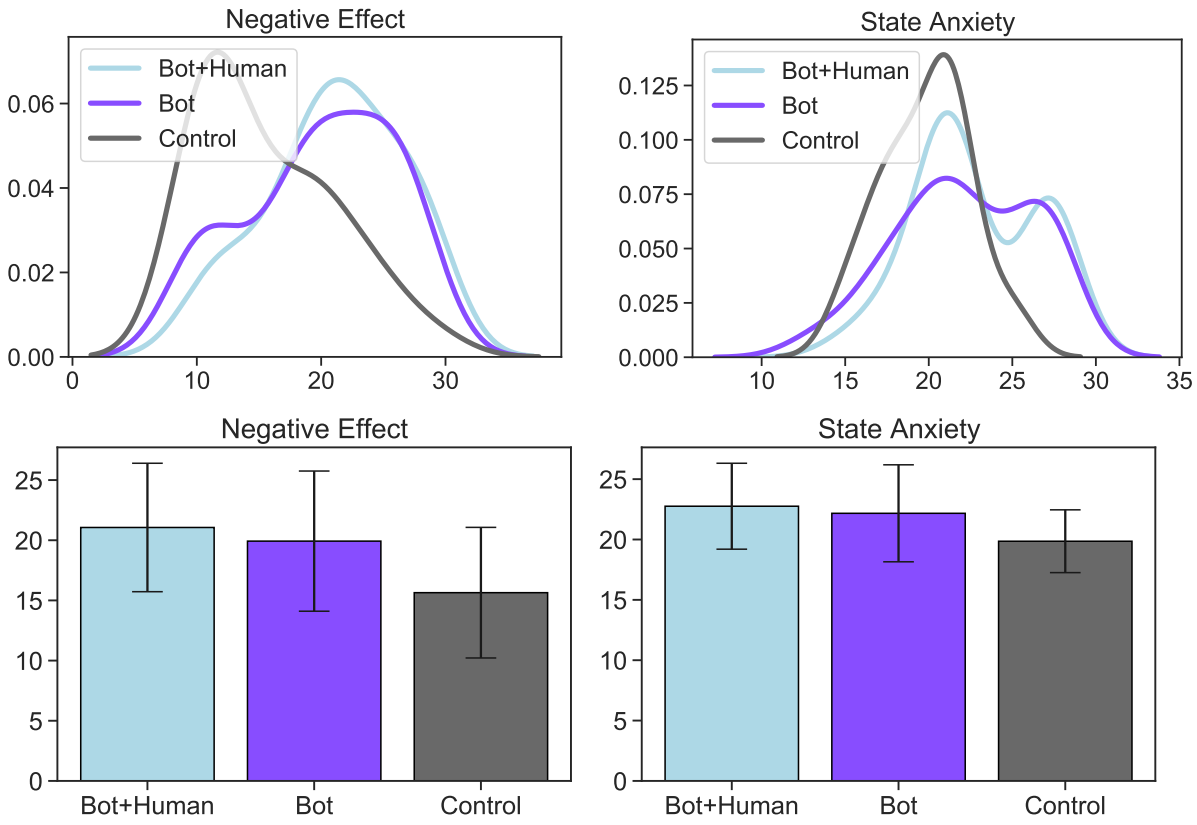

Figure 4: Validation experimental results. Bot and Bot+Human stories significantly amplify negative affect and increase state anxiety compared to the control group.

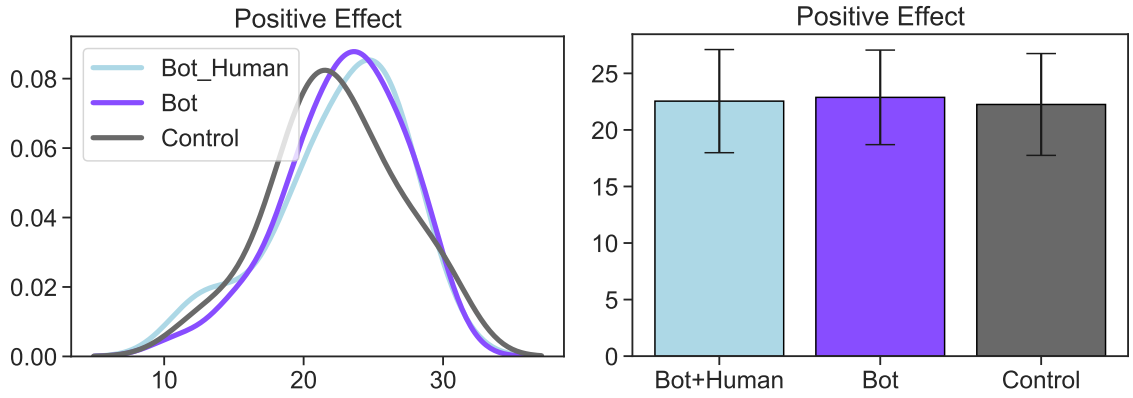

Figure 5: Validation experimental results. Bot and Bot+Human conditions reduce positive affect as compared to the control group, albeit without a significant difference.

share 1) a large-scale horror story dataset collected from Reddit's r/nosleep subreddit, 2) a labeled dataset on generated stories indicating how scary they are on a scale of 1-5, 3) a tree-based dataset of 500 stories created as a collaboration between our bot and Twitter users. We believe these datasets will encourage further research on this topic and answer several important research questions. As future work, our approach can be to improve the performance of the story generation system by tailoring the preferences towards particular users or can be explored to understand what particular features of the generated stories induce fear.

REFERENCES

[1] [n.d.]. *AI art: I'm using a language model called GPT-2 to write my next novel - Vox.* https://www.vox.com/future-perfect/2019/8/30/20840194/ai-art-fiction-writing-language-gpt-2

[2] [n.d.]. *Botnik – Human-machine entertainment.* https://botnik.org/

[3] [n.d.]. *Deep Dream Generator.* https://deepdreamgenerator.com/

[4] [n.d.]. *deepart.io - become a digital artist.* https://deepart.io/

[5] [n.d.]. *Magenta.* https://magenta.tensorflow.org/

[6] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. *arXiv preprint arXiv:2010.01717* (2020).

[7] Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. 2020. Cue Me In: Content-Inducing Approaches to Interactive Story Generation. *arXiv preprint arXiv:2010.09935* (2020).

[8] Gwern Branwen. [n.d.]. *GPT-2 Neural Network Poetry · Gwern.net.* https://www.gwern.net/GPT-2

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[10] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. [n.d.]. How Novelists Use Generative Language Models: An Exploratory User Study. ([n. d.]).

[11] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y Lim. 2021. Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation. *arXiv preprint arXiv:2101.06030* (2021).

[12] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[13] Pablo Gervás, Raquel Hervás, and Carlos León. 2015. Generating Plots for a Given Query Using a Case-Base of Narrative Schemas.. In *ICCBR (Workshops)*. 103–112.

[14] Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6473–6480.

[15] Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. *arXiv preprint arXiv:1707.05501* (2017).

[16] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[17] Ahmed Khalifa, Gabriella AB Barros, and Julian Togelius. 2017. Deeptingle. *arXiv preprint arXiv:1705.03557* (2017).

[18] Joy Kim, Justin Cheng, and Michael S Bernstein. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 745–755.

[19] Carolyn Lamb. 2017. Crowdsourced Social Media Poetry. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 13.

[20] Michael Lebowitz. 1987. Planning stories. In *Proceedings of the 9th annual conference of the cognitive science society*. 234–242.

[21] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. 2018. Diverse Image-to-Image Translation via Disentangled Representations. *CoRR* abs/1808.00948 (2018). arXiv:1808.00948 http://arxiv.org/abs/1808.00948

[22] Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27.

[23] Hugo Liu and Push Singh. 2002. MAKEBELIEVE: Using commonsense knowledge to generate stories. In *AAAI/IAAI*. 957–958.

[24] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101* (2016).

[25] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP*.

[26] Theresa M Marteau and Hilary Bekker. 1992. The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology* 31, 3 (1992), 301–306.

[27] James R Meehan. 1977. TALE-SPIN, An Interactive Program that Writes Stories.. In *Ijcai*, Vol. 77. 9198.

[28] Rutherford MONTGOMERY. 2009. The Abominable Snowman (Choose Your Own Adventure# 1). *Choose Your Own Adventure Series.[Kindle Version] Chooseco* (2009).

[29] Hugo Gonçalo Oliveira. 2017. O Poeta Artificial 2.0: Increasing meaningfulness in a poetry generation twitter bot. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. 11–20.

[30] D Pawade, A Sakhapara, M Jain, N Jain, and K Gada. 2018. Story scrambler-automatic text generation using word level RNN-LSTM. *International Journal of Information Technology and Computer Science (IJITCS)* 10, 6 (2018), 44–53.

[31] Project Gutenberg. [n.d.]. .

[32] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[34] Mark Owen Riedl and Vadim Bulitko. 2013. Interactive narrative: An intelligent systems approach. *Ai Magazine* 34, 1 (2013), 67–67.

[35] Melissa Roemmele and Andrew S Gordon. 2015. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*. Springer, 81–92.

[36] Melissa Roemmele, Sosuke Kobayashi, Naoya Inoue, and Andrew Gordon. 2017. An RNN-based binary classifier for the story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. 74–80.

[37] Arthur L Samuel. 1967. Some studies in machine learning using the game of checkers. II—Recent progress. *IBM Journal of research and development* 11, 6 (1967), 601–617.

[38] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[39] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* (2017).

[40] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2017. Cold fusion: Training seq2seq models together with language models. *arXiv preprint arXiv:1708.06426* (2017).

[41] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *ArXiv* abs/1409.3215 (2014).

[42] Reid Swanson and Andrew S Gordon. 2008. Say anything: A massively collaborative open domain story writing companion. In *Joint International Conference on Interactive Digital Storytelling*. Springer, 32–40.

[43] Edmund R Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of cross-cultural psychology* 38, 2 (2007), 227–242.

[44] Tony Veale and Mike Cook. 2018. *Twitterbots: Making Machines that make meaning*. MIT Press.

[45] Tony Veale, Alessandro Valitutti, and Guofu Li. 2015. Twitter: The best of bot worlds for automated wit. In *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 689–699.

[46] Ben Verhoeven, Iván Guerrero, Francesco Barbieri, Pedro Martins, and Rafael Pérez y Pérez. 2015. TheRiddlerBot! (2015).

[47] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* (2017).

[48] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[49] Wikipedia. 2021. Darkness (poem) — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Darkness%20(poem)&oldid=967764144. [Online; accessed 15-February-2021].

[50] Wikipedia. 2021. Fragment of a Novel — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Fragment%20of%20a%20Novel&oldid=987555806. [Online; accessed 15-February-2021].

[51] Wikipedia. 2021. Frankenstein — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Frankenstein&oldid=1006555968. [Online; accessed 15-February-2021].

[52] Wikipedia. 2021. The Vampyre — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=The%20Vampyre&oldid=1003973483. [Online; accessed 15-February-2021].

[53] Wikipedia. 2021. Year Without a Summer — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Year%20Without%20a%20Summer&oldid=1004991548. [Online; accessed 15-February-2021].