Perceptions of Human and Machine-Generated Articles

SHUBHRA TEWARI, Whitman College
RENOS ZABOUNIDIS, University of Massachusetts Amherst
AMMINA KOTHARI, REYNOLD BAILEY, and CECILIA OVESDOTTER ALM, Rochester Institute of Technology

Automated journalism technology is transforming news production and changing how audiences perceive the news. As automated text-generation models advance, it is important to understand how readers perceive human-written and machine-generated content. This study used OpenAI's GPT-2 text-generation model (May 2019 release) and articles from news organizations across the political spectrum to study participants' reactions to human- and machine-generated articles. As participants read the articles, we collected their facial expression and galvanic skin response (GSR) data together with self-reported perceptions of article source and content credibility. We also asked participants to identify their political affinity and assess the articles' political tone to gain insight into the relationship between political leaning and article perception. Our results indicate that the May 2019 release of OpenAI's GPT-2 model generated articles that were misidentified as written by a human close to half the time, while human-written articles were identified correctly as written by a human about 70 percent of the time.

CCS Concepts: • **Human-centered computing** → *Laboratory experiments*; • **Computing methodologies** → *Natural language generation*; • **Social and professional topics** → Cultural characteristics;

Additional Key Words and Phrases: Automated text generation, sensing readers, facial expressions, galvanic skin response, political topics, political leaning

ACM Reference format:
Shubhra Tewari, Renos Zabounidis, Ammina Kothari, Reynold Bailey, and Cecilia Ovesdotter Alm. 2021. Perceptions of Human and Machine-Generated Articles. *Digit. Threat.: Res. Pract.* 2, 2, Article 12 (April 2021), 16 pages.
https://doi.org/10.1145/3428158

1 INTRODUCTION

The boundaries between journalism, automation, and machine-generation of content are increasingly being blurred. Many news publishers have embraced computational journalism as a strategy to generate content in an efficient and cost-effective manner [38], with varying opinions on what the ideal relationship between news

S. Tewari and R. Zabounidis contributed equally to this research.

This material is based upon work supported by the National Science Foundation under Award No. IIS-1851591. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Authors' addresses: S. Tewari, Whitman College, 345 Boyer Ave, Walla Walla, WA, 99362; email: tewaris@whitman.edu; R. Zabounidis, University of Massachusetts Amherst, 120 Tillson Farm Road, Amherst, MA 01003; email: rzabounidis@cs.umass.edu; A. Kothari, R. Bailey, and C. O. Alm, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY, 14623; emails: abkgpt@rit.edu, rjb@cs.rit.edu, coagla@rit.edu.

2576-5337/2021/04-ART12
https://doi.org/10.1145/3428158

automation algorithms and journalistic content should be [4]. Computer-assisted reporting, which involves the use of computers to analyze data, and computational journalism, which includes knowledge representation using AI and computational tools, are both forms this relationship can take [9]. A recent survey of news organizations globally found that artificial intelligence has become a significant part of news gathering, production, and distribution processes [3]. While the technological advances promise to extend journalists' ability to produce more content faster, concerns about natural language processing models generating news texts bring up ethical and editorial issues. With the increasing use of AI and automation in newsrooms, news content generated by software is becoming more difficult to separate from articles written by human journalists [8]. In addition, with readers shifting to online journalism and social media, the chances of them being exposed to misinformation and fake news have increased. Furthermore, the rapid spread of online news has resulted in more readers being exposed to inaccurate content that is passing as credible information. We thus seek to test the ability of a natural language processing model to mimic human-written texts by studying readers' detection and perceptions of human-written and machine-generated stories.

Some researchers have examined users' perceptions of machine-generated articles when compared to articles written by human journalists [8, 16, 37]. Our study builds upon this work as well as the work of those who explored the influence of political leaning on how readers respond to machine-generated news about controversial topics [20, 21, 38]. Some readers perceive online news to be more credible than print news [5], and there is an increasing possibility that readers online come across written content auto-generated by algorithms. With the development of continuously more sophisticated text-generation models, which might impact future experiences and manipulation of readers, it is important to explore the credibility of machine-generated text content.

OpenAI developed a text generation model, called GPT-2, to generate articles based on a prompt given by the user [30]. However, they deemed this model too risky to be released to the public due to concerns about malevolent use. They released a restricted version of the model in May 2019. Extended models have since been released. In this study, 30 participants were presented with a randomized selection of human- and machine-generated articles (validated for syntactic and semantic fluency), while monitoring their facial expressions and galvanic skin response. We examined OpenAI's claim that GPT-2 could closely mimic human-written text as a journalistic form of the Turing Test, and considered the relationship between political leaning and readers' perceptions of articles.

2 RELATED WORK

2.1 Information Credibility

In an era of misinformation, the importance of determining information credibility becomes ever more apparent. A study on perceptions of information credibility related to print and online news sources found that internet news content was considered moderately credible overall, and when digital journalists were asked their preferences, they choose online news information [5]. Further research on perceptions of news credibility, specifically in the context of the Iraq war, showed that opponents of the war felt that the internet was a more credible source of news than those who were neutral or in support of the war [7]. However, another study has found that people are less skeptical of print news compared to online and television news sources [17].

A model to analyze how people perceive the credibility of information provided by technology suggests that readers' interactions via the modality, agency, interactivity, and navigability of the technology contribute to their overall perception of its credibility [34]. Studies of the methods used to characterize and detect disinformation suggest that data mining approaches have potential to improve the detection of fake news [32], and have also identified relevant factors such as the media outlets publishing the news, social media users who share the news, and the tone of the news stories [26]. Research is also being conducted to explore questions beyond the detection of fake news such as measuring its impact on society [27].

2.2 Text Generation

In addition to OpenAI, many other organizations have developed methods of automated text generation, such as Narrative Science and CBS Interactive [37]. Advanced neural network approaches have eclipsed earlier models and been applied to a range of text generation tasks and corpora. Both supervised and unsupervised models have been utilized to characterize the political leaning of various news sources [29], and prediction of trustworthiness and political leaning of various news outlets has also been considered [1]. Models such as Proppy, for example, explore levels of propaganda in news articles [2]. Efforts to determine whether news sources and articles are based on credible information also motivate studying reader perceptions of machine-written content. Another concern is the risk of ethically problematic applications being developed and applied to news-like content [14].

Both supervised and unsupervised models have been utilized to classify the political leaning of various news sources [29], and prediction of trustworthiness and political leaning of various news outlets has also been explored [1]. Models such as Proppy, for example, aim to identify levels of propaganda in news articles [2], and this establishes a foundation for methods used to determine whether news sources and articles are based on credible information. Moreover, there is an on-going risk that ethically problematic applications are developed and applied to news-like content [14].

2.3 Readers' Perceptions Based on Source

A previous study by Clerwall [8] exploring perceptions of automatically generated content versus articles written by journalists presented readers with an auto-generated recap of a Los Angeles Chargers game and a human-written article about the NFL from the LA Times to see if they can tell the difference. The study found that readers were not able to make a clear distinction between these articles [8]. Graefe et al. [13] noted that readers found machine-generated sports articles to be slightly more credible than the articles written by human journalists.

Furthermore, when journalists and general readers were shown human- and software-generated baseball game articles, both responded more positively to articles when told they are machine-authored than when told they were human-authored [16]. However, it is not conclusive that people trust machine-generated articles more than human-written content. For example, a Dutch study asking both journalists and general readers of news to rank articles about finance and sports on expertise and trustworthiness constructs found that general participants ranked content written by software and content written by a journalist to have equal levels of trustworthiness and expertise [37]. The study also showed that journalists ranked software-generated articles as having higher expertise but human-written articles higher for trustworthiness.

Research has also demonstrated that many psychological factors influence how people interact with technology. For example, research into the tendency to distinguish between humans and computers, involving subjects over the age of 20, found that subjects were uncomfortable with computers in personal roles [23]. Further explorations into participants' responses to a virtual reality chat found that participants felt more unsettled when they perceived an avatar to be controlled by an artificially intelligent computer [33]. While previous studies on readers' perceptions of news content have focused on articles covering empirical data related to formulaic topical content, readers' perceptions of politically charged articles is understudied. The OpenAI GPT-2 model we explore in this study generates more flexible open-domain content.

2.4 Research Questions

In this study, we apply GPT-2 conditional text generation to examine readers' assessment and reactions to machine-generated political content, focusing on topics that engage readers' emotions. We define *human-written* as news content written by human journalists, while *machine-generated* is defined as articles that have been automatically generated by an AI system. *Political leaning* is the term used for the participants' own self-identified political leaning on a Liberal-to-Conservative spectrum. Articles were also assigned a liberal or conservative

category by two co-authors.[1] In addition, *political tone* refers to the Right-to-Left leaning tone that the text conveyed to the participant. Last, perceived *credibility* of an article is the perceived trustworthiness of its content. We answer the following questions:

(1) Are participants able to identify the source of an article: human or machine?
(2) Are there differences in human reactions to:
    (a) human-written versus machine-generated articles?
    (b) conservative versus liberal articles?
(3) Does readers' political leaning impact perceived credibility of articles?

3 METHODS

3.1 Article Generation and Selection

Machine-generated articles were prepared using the approximately 350 million parameter version of the GPT-2 model, which took an input text and returned a roughly 400-word article. If a chosen article was longer than 400 words, roughly the first 400 words were selected, allowing subjects to read the articles within a reasonable amount of time. The given prompts were either conservative or liberal on one of three controversial topics: climate change, vaccinations, or politics. For each category of articles (e.g., conservative on climate change, liberal on vaccinations, etc.), we prompted the article generator with the first few sentences from a human-written article of the same category (e.g., the first two sentences of a human-written conservative article on climate change used to prompt a machine-generated conservative article on climate change). We observed that GPT-2 often strayed from the given topic and/or the expected political leaning. Accordingly, for each prompt, 20 samples were generated and the one that was deemed most coherent, on-topic, and with desired political leaning based on consensus of two experimenters was chosen. This practice is consistent with how OpenAI developers generated their example texts [30].

Human-written articles were selected from well-known American news publications from a range of political affiliations. These news sources were from a dataset listing American publications with known political leanings [35]. Articles were chosen based on political affiliation and topic with the goal to obtain a balanced number of liberal and conservative articles, and a balanced number of articles covering climate change, vaccinations, and politics. Selected articles were shortened to around 400 words to match the length of the machine-generated articles. In total, 48 articles were selected for the pool of articles from which 12 were then randomly presented to each participant.

3.2 Journalistic Turing Test

Turing formally defined a test aimed to determine whether a computer could act indistinguishably from a human. Since then, mention of the "Turing Test" has evolved to refer to a wide range of behavioral tests [24, 36]. We define a *journalistic Turing Test* to measure the ability of a machine to generate news-like articles that are indistinguishable from human journalistic writing on the same topic. In this study, it is important to note that participants were not experts, but regular readers who did not necessarily have a high media literacy. This differentiates the journalistic Turing Test from the from an expertise-based test like the Feigenbaum Test [12], where the participants would be subject experts who are more likely able to identify the difference between a human and an AI system.

3.3 Peak Expression Analysis

The study collected and analyzed human sensing data from participants reading articles, including facial expressions and galvanic skin response (GSR). For the data processing, standalone dimensionality reduction algorithms

[1]Initial capitalization of these terms is used to indicate the self-identified political leaning.

Fig. 1. Top: Setup of articles and surveys in experiment. Bottom: Webcam facial expressions and GSR visualization using iMotions.

were not directly applicable due to the fact that GSR and facial data time series are variable in length, since readers read at different speeds. In addition, most readers may not exhibit emotional reactions for the majority of the time reading, so some parts of the time series are more important than others. To account for this, analysis first detected peaks in GSR data. GSR peaks tend to relate to affective peaks. The analysis then considered face-based emotional expressions during peaks [10]. Averaging each expression resulted in the time series being reduced to a single n-dimensional vector, where n was the number of expressions. In order to visualize results, we further reduced this n-dimensional vector down to two dimensions using Principal Component Analysis [19]. We refer to this process as Peak Expression Analysis (PEA).

3.4 Experiment

3.4.1 Participants. In total, 30 participants were recruited through flyers and emails after receiving Institutional Review Board approval. Participants were chosen on a first-come, first-serve basis while ensuring an even distribution of male and female participants. Participants were reimbursed with $10 for their time. While our sample size is modest, given the exploratory focus of our study, we provide a methodology and findings that still provide valuable bench-marking data and can later be considered for similar work with larger groups of participants.

3.4.2 Experimental Setup. Before the experiment, each participant was connected to a Shimmer GSR sensor on their pointer and middle fingers on their non-dominant hand (see Figure 1, bottom-left), to monitor pulse and skin conductance as they participated in the experiment. Each participant was then asked to complete a demographic survey, providing information about their gender, race, age, education level, employment status, and political identification. They were also asked to rank how strongly they agreed with statements on climate change, vaccinations, and political issues (see Appendix A). In the experiment, individual participants were presented with 12 randomly selected article excerpts (half from each source category and further split based on the

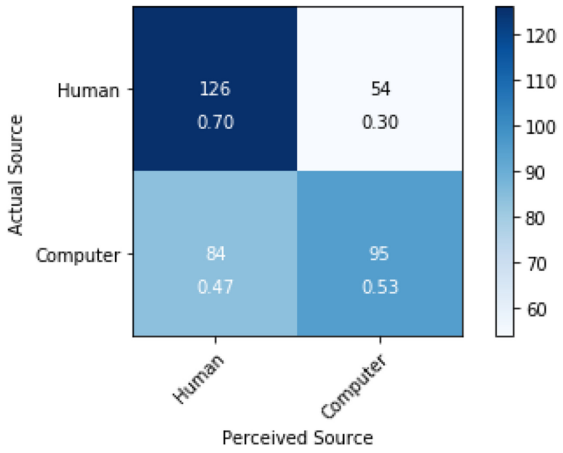

Fig. 2. Actual versus perceived source (N = 359 data points).

three topics) and after each reading task they were presented with a questionnaire asking them to summarize the article, how credible they thought the article was, whether they believed the article was written by a human or by a computer, how engaging the article was,[2] and what the political leaning of the article was (see Appendix B). One item was excluded from analysis since this questionnaire was not completed for it.

Our study focused on the ability of human readers to discern differences between human- and machine-generated articles and evaluate the credibility of the information. Prior research shows that media message credibility can be affected by numerous factors making it difficult to determine if the audience's perception of the credibility is based on the source, medium, or the message [22]. Hence, our study employed a single-item credibility measure to assess if the participants found the articles credible based on reading them. In order to gather biophysical data to gauge participants' responses to articles beyond the self-reported survey data, facial expressions were recorded through the computer's web camera and their facial movements were analyzed through iMotions Facial Expression Analysis software, which provided face-based estimations on each participants' levels of valence, engagement, joy, anger, surprise, fear, contempt, sadness, and disgust while reading.

Participants completed a post-experiment questionnaire asking them about where they get their news from, how often they consume news content, what their preferred news organization was, based on a list of popular news publishers (ranging in political leaning), if they enjoy discussing news, if they knew that some news sources produce machine-generated articles, how comfortable they were with an algorithm customizing news stories for them based on data gathered from their online activity, and if they would prefer a story written by a human or by computer software.

4 RESULTS

4.1 RQ1: Are Participants Able to Identify the Source of an Article: Human or Machine?

A Chi-Square test of independence was conducted to determine if there was a significant relationship between two categorical variables: actual versus perceived source (human or computer). The relationship between these variables was found to be significant, χ^2 (1, N = 359) = 19.679, p< 0.01. Figure 2 shows the number of times people perceived a human-written article was written by a human versus a machine and how often people thought a machine-generated article was written by a human or by a machine. The machine-generated articles were perceived to be written by a human 47% of the time, while the human-written articles were identified as

[2]Participants were notified of a wording issue in the engagement question prior to completing the task.

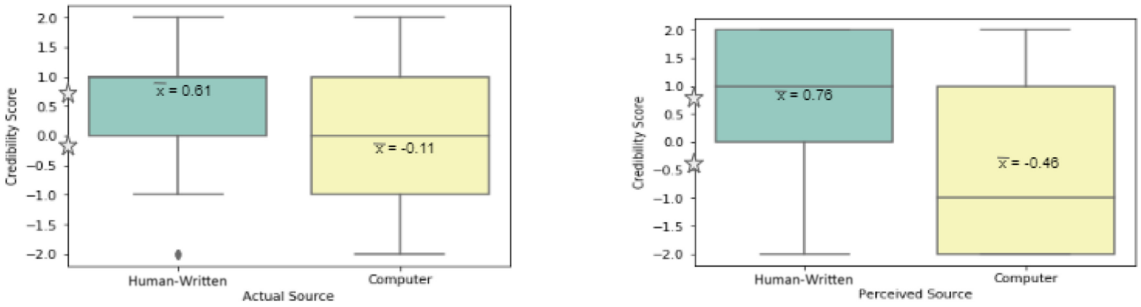

Fig. 3. Left: Credibility scores for human- versus machine-generated articles. Right: Credibility scores for articles perceived to be human- versus machine-generated. Stars indicate the means.

being written by a human about 70% of the time. This suggests that people can identify human-written articles with relatively more confidence than machine-generated articles, based on our sample.

4.2 RQ2a: Are There Differences in Human Reactions to Human-Written Versus Machine-Generated Articles?

An independent samples Mann-Whitney U test was conducted to compare differences between the credibility score, on a scale from −2 (very uncredible) to 2 (very credible), given to human- versus machine-generated articles. Credibility was treated as an interval measure in which each of the five options given to participants (very uncredible, somewhat uncredible, neutral, somewhat credible, very credible) were associated with a number between −2 and 2. The sample articles in our experiment did not indicate the name of the source as our interest was in accessing general perception based on the content of the article and whether the participants could differentiate between human- and machine-generated content. This test was used because there were two independent groups (human- and machine-generated articles) with a dependent variable (credibility score) that was not normally distributed (Shapiro-Wilk Test, $p < 0.05$). Machine-generated texts had significantly lower credibility scores (Mdn = 0.00) than human-written texts (Mdn = 1.00), $U = 11473$, $p < 0.01$ (Figure 3, left). Follow-up independent samples Mann-Whitney U test indicated that, based on participants' perceptions of articles' source, texts that were perceived to be machine-generated also received significantly lower credibility scores (Mdn = −1.00) than those that were perceived to be human-written (Mdn = 1.00), $U = 7568$, $p < 0.01$ (Figure 3, right).

Next, the peaks in GSR response for each participant were used to conduct an independent samples Mann-Whitney U test to determine if there was a significant difference between the mean number of peaks that occurred while participants read human-written articles versus those that occurred while they read machine-generated articles (Figure 4, left and right). Again, this test was chosen because there were two independent groups (human-written and machine-generated) and a continuous dependent variable (mean GSR peak count). No significant differences were found.

During GSR peaks [11], participants' facial expressions were analyzed using PEA to visualize facial and GSR time series data when reading human-written and machine-generated articles (Figure 5). As a first step, Principal Component Analysis (PCA) was conducted with scikit-learn [28], using the nine face-based estimated expressions as input features derived using PEA with the dataset. The clusters formed suggest a potential link between facial expression and credibility.

4.3 RQ2b: Are There Differences in Human Reactions to Conservative Versus Liberal Articles?

To determine if there was a relationship between participants' GSR responses to conservative and liberal articles and self-identified political affiliation (grouping participants who identified as *Right* or *Center Right* as

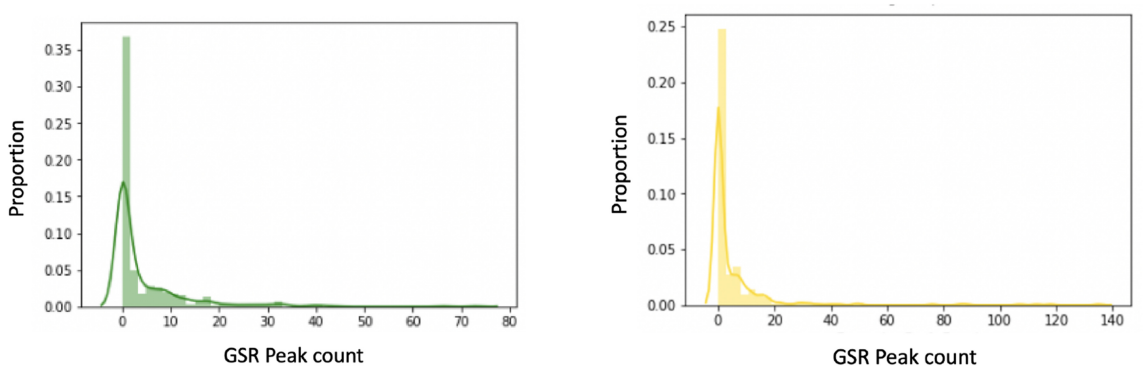

Fig. 4. Left: Distribution of GSR peaks when reading human-written articles. Right: Distribution of GSR peaks when reading machine-generated articles.

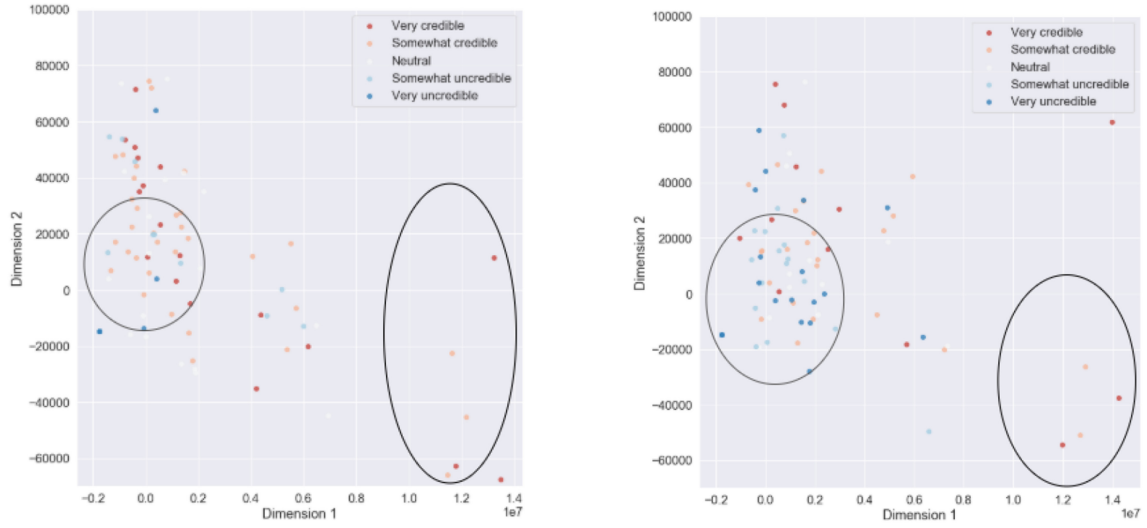

Fig. 5. Left: Clusters of facial expressions based on credibility rating given to human-written articles during GSR peaks. Right: Clusters of facial expressions based on credibility rating given to machine-generated articles during GSR peaks (N = 359 data points).

Conservative and participants who identified as *Far Left*, *Left*, or *Center Left* as *Liberal*; see Figure 6), an independent samples Kruskal-Wallis test was conducted ($H(2) = 8.706$, $p = 0.013$). The test showed a significant difference between the number of GSR peaks for Liberals, with a mean score of 5.10, Independents, with a mean score of 4.07, and Conservatives, with a mean score of 4.83, while reading liberal articles. This test was chosen because we had multiple independent groups (Liberal, Independent, Conservative) and a continuous dependent variable (mean GSR peak count). No significant difference was found between GSR peaks of these three groups while participants were reading conservative articles.

Participants' political leaning was used as an independent variable. Our sample was skewed towards participants who identified as Liberal or Independent. Other demographic factors such as age and occupation were too homogeneous among the modest sample, since our experiment was conducted on a university campus, to be able to conduct tests considering these factors. During GSR peaks, facial expressions were also analyzed using

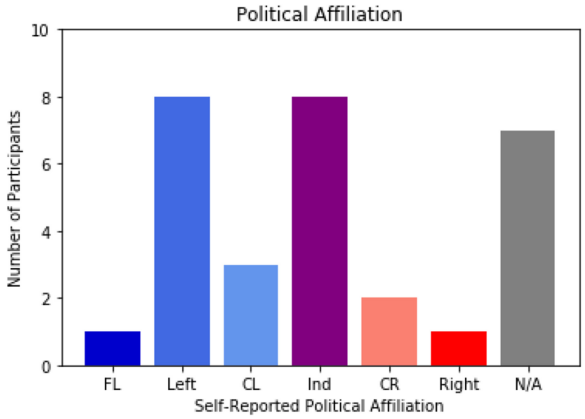

Fig. 6. Distribution of self-reported political leaning. N/A indicates participants who chose not to answer. No one indicated Far Right (FR).

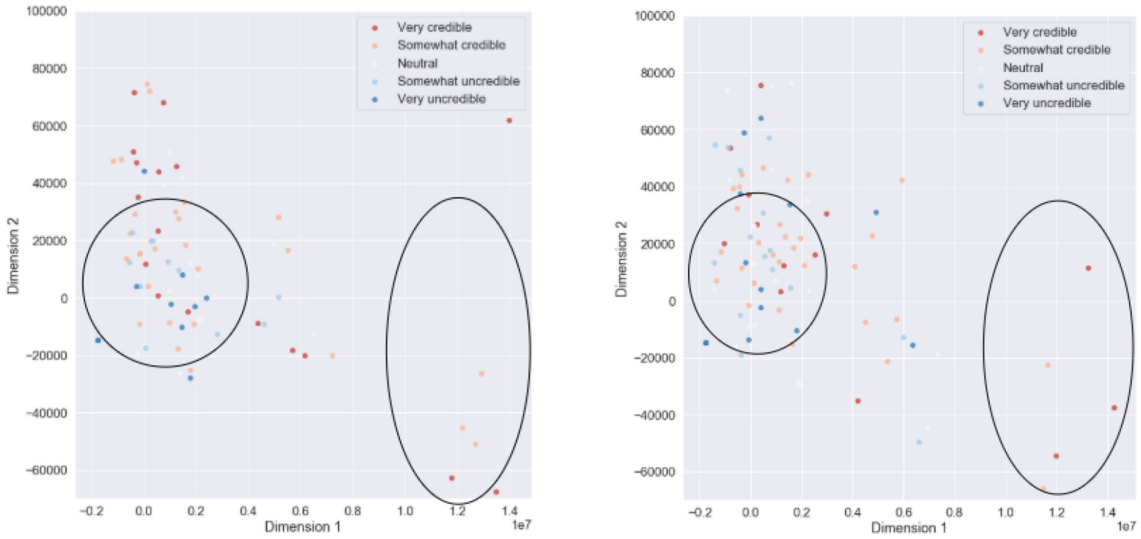

Fig. 7. Left: Clusters of facial expressions based on credibility rating given to liberal articles during GSR peaks. Right: Clusters of facial expressions based on credibility rating given to conservative articles during GSR peaks.

PEA for participants reading both liberal and conservative articles (Figure 7). The groupings weakly indicate that facial expressions of readers around the moments of GSR peaks can give insight as to how readers rate the credibility of articles.

4.4 RQ3: Does Readers' Political Leaning Impact Perceived Credibility of Articles?

An independent samples Mann-Whitney U test was conducted to compare differences between two independent groups (Liberals and Conservatives; see Section 4.3) with a dependent variable (credibility score) that was not normally distributed (Shapiro-Wilk test, $p < 0.05$). Results indicated that there was not a significant difference between credibility scores given to human-written conservative, machine-generated conservative, human-written liberal, and machine-generated liberal articles by Liberal versus Conservative participants.

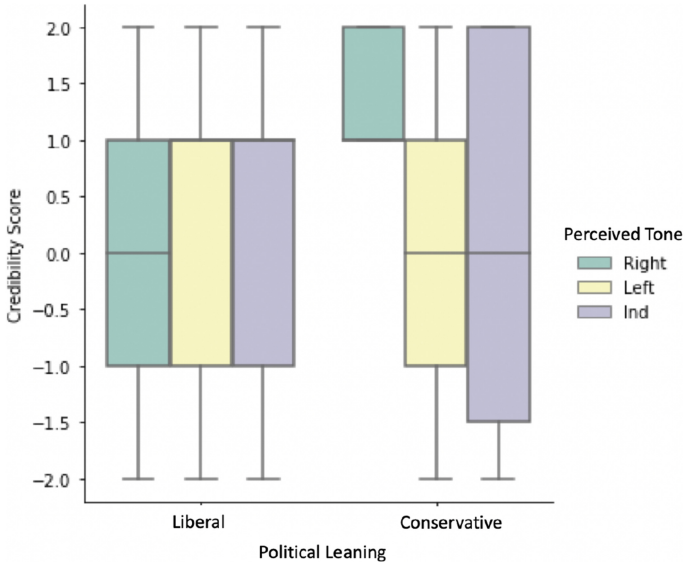

Fig. 8. Distribution of credibility scores for articles perceived to have a Left, Right, or Independent tone.

Since articles were stratified by three topics—climate change, vaccinations, and political news—each participant was asked to provide their opinion on these three issues in the demographic survey. Independent samples Mann Whitney U tests were conducted to determine the relationship between participants' opinions on each of these issues and their perceived credibility of articles on each respective issue. All participants reported that they either *Strongly agree* or *Somewhat agree* that climate change is an important issue, and there was no significant difference found between the credibility scores provided by these two groups on climate change articles. Similarly, all participants reported that they either *Strongly agree* or *Somewhat agree* that everyone should get their children vaccinated and there was no significant difference found between these two groups' credibility scores for articles on vaccinations. Finally, participants who responded that they either *Strongly agree* or *Somewhat agree* that Donald Trump is a good president were placed in an *Agree* group and participants who chose *Strongly disagree* or *Somewhat disagree* were placed in a *Disagree* group. There was no significant difference between credibility scores given to political articles by these two groups.

Lastly, an independent samples Mann-Whitney U test was conducted based on the same two Liberal and Conservative groups' credibility scores given to articles perceived as either *Far Left*, *Left*, or *Center Left* in one group as *Left*, *Independent* in a second group, and *Far Right*, *Right*, or *Center Right* in a third group as *Right* (Figure 8). For articles perceived to have a Right-leaning tone, credibility scores given by Liberals were significantly lower (Mdn = 0.00) than those given by Conservatives (Mdn = 1.00), $U = 240.5$, $p = 0.001$. A significant difference was not found for articles perceived as having Left or Independent political tones.

5 DISCUSSION

The goal of our study was to examine if regular news readers can differentiate between machine-generated and human-written articles. Our findings show that the machine-generated articles were identified as human roughly half the time, while human-written articles were identified correctly as written by a human about 70 percent of the time. Since people are able to identify the article source better for human-written articles than for machine-generated articles, this finding differs from Clerwall's conclusion that readers are not able to make a clear distinction between software- and human-written articles [8]. However, Clerwall's study used

articles focusing on American football game scores, which tend to be formulaic in nature, whereas the present study's articles were more politically charged. In addition, while Clerwall used formulaic models of text generation, GPT-2 used template-free, neural-based generation. As a result, the GPT-2 model sometimes generated text that might be semantically and syntactically coherent, but whose topic could change in a way humans find unnatural. It is important to note that participants in our study were aware that some of the articles they were reading were machine-generated, and were thus primed to detect such flaws.

Participants in this study demonstrated skepticism of machine-generated content. When exploring to see if there were differences in reactions to human-written versus machine-generated articles, we found that machine-generated articles had significantly lower credibility scores than human-written articles. We further tested credibility score differences between articles that were perceived to be human-written versus machine-generated. Perceived machine-generated articles had significantly lower credibility scores based on this test as well. It is important to note that if people know that there is a possibility of an article being machine-generated, they might be less likely to believe its content. Our finding varies from Jung et al. [16]'s conclusion that journalists and the general public rate article content identified as machine-generated higher than that identified as journalist-written. However, they utilized baseball game articles for their study as well as different text-generation technologies than our study. While our results are reassuring given concerns about fake news and limitations of natural language generating models, the possibility of misleading machine-generated text being perceived as credible remains.

Looking at the GSR data collected to answer whether there were differences in reactions to human- versus machine-generated articles, we found a significant difference between the number of peaks for participants leaning Liberal or Conservative or identifying as Independents when reading liberal articles. The trend indicated that Liberals were more engaged with liberal news content, which is somewhat surprising since one might assume that Conservatives would become agitated when reading liberal content, although we do not know if GSR peaks indicate positive or negative engagement. Measuring arousal via GSR while reading human- and machine-generated texts is a unique contribution of this study, as previous works have focused on self-reported survey data.

6 CONCLUSION

The advancement of automated journalism has the potential to significantly change the news content publishers produce and readers consume on a daily basis, but greater attention needs to be paid to how AI and machine-learning methods are integrated due to on-going ethical and editorial concerns about inaccuracies and bias. This work contributes to scholarship on the development of a methodology for assessing readers' reactions to machine-generated news articles. Our study has also developed a dataset consisting of 359 instances of facial expressions, GSR, and self-reported perceptions of 24 human-written and 24 machine-generated articles, which can be used for future research. The application of a journalistic approximation of a Turing Test, and analysis considering credibility of textual content, the source, political leaning of participants and content, and perception of textual content's political tone on a carefully collected dataset provides a good benchmark to develop further research on this topic. Results suggest that readers can still distinguish between human-written and machine-generated articles and can perceive the lack of credible information in machine-generated content. This suggests that humans can exhibit critical resilience to machine-generated fake news and highlights the current limitations of this natural language processing model to successfully mimic human text generation. Yet, the ethical issues with news-like story generation remain, and there is a need for critical awareness not only among readers but also among researchers, professional software developers, and data scientists, considering the potential for improved capacity of text-generation models.

In this study, we have reported on responses to machine-generated articles using both readers' self-reported perceptions and reactions such as GSR and facial expressions. Additionally, based on the collected dataset, we have found that articles generated by OpenAI's approximately 350 million parameter GPT-2 model were identified as human-written roughly half the time. However, the findings do not preclude the potential that models will

one day approximate human-written content. In this study, readers were more confident about human-written articles, identifying their source correctly about 70 percent of the time. It also remains to be seen how people would perceive articles generated by more recent versions of GPT-2 with substantially more parameters.

The data analysis also suggests that readers do not yet trust software to write credible content for politically charged topics. While a comparison remains to be done between formulaic articles on sports or finance and controversial articles such as the ones presented in our study, when compared to Jung et al. [16], who used sports articles in their study, the present study resulted in significantly lower credibility scores for machine-generated articles than for human-written text.

Finally, based on participants' self-reported political affiliations, results indicate that Liberals ranked articles perceived to have a Right-leaning tone to be significantly less credible than Conservatives did. We also found that the Liberal readers had more GSR peaks when reading liberal articles. These findings point to higher engagement of Liberal readers when reading content that they presumably agree with (or at least do not strongly disagree with) while also lower levels of trust for content they presumably do not agree with. Once again, future research should expand upon this dataset and these results to confirm the relationship between participants' political self-identification and their response to news-like content.

6.1 Limitations and Future Work

While our study makes contributions to scholarship on news credibility assessment and natural language processing models' ability to mimic human-written text, our sample was heavily skewed towards university students and people who identified as either Liberal or Independent. This means that we had little representation of politically Conservative readers, which might have impacted our results related to the relationship between political leaning and perceptions of news credibility.

We also used a single-item measure of credibility, which is not the most reliable choice of measurement, but given the long duration of our experiment and lack of information about news publications, author byline, dateline, or link to other sources in the story, it was sufficient for our study. Future work should incorporate some of the existing news credibility measures to test whether perceptions of credibility change the author's byline or if links to other sources make a difference in how people perceive human-written versus machine-generated stories.

Future studies could additionally include more participants and further break down the articles presented to participants by different genres. Researchers could also consider using more information about the provided articles, such as date, as further measures of news credibility. Since stories on finances and sports are more formulaic in nature because they are based on scores, numbers, and statistics, one might expect such machine-generated text to seem more realistic compared to more politically charged topics. One approach would be breaking up participants into two groups: one group that reads political and controversial articles as we did in this experiment, and one group that reads more structured articles on finances or sports statistics. Furthermore, to better understand whether participants' facial expressions and GSR response were actually in response to the articles or were just their ordinary facial expressions and GSR, neutral articles, or affective palate cleanser tasks [31], could be interspersed between the different reading items to compare how participants respond to non-provocative content. In addition, as our study was skewed towards Liberal and Independent participants with little representation of Conservatives, future work could seek a more diverse range of political leanings. Based on our study, and its findings for the journalistic Turing Test, more work is needed to test the ability of text generation models to mimic human-written content.

APPENDICES

A PARTICIPANT DEMOGRAPHICS SURVEY

(1) Subject ID:

(2) Gender:
- Female
- Male
- Prefer not to answer
- Other

(3) Race:
- White (Not Hispanic or Latino)
- Asian/Pacific Islander
- Hispanic or Latino
- Black or African American
- Native American or American Indian
- Prefer not to answer
- Other

(4) Age:
- Under 24
- 25–35
- 36–45
- 46–55
- 56–65
- Over 65

(5) Education Level:
- PhD
- Masters
- Bachelors
- High School Diploma
- No Degree
- Other

(6) Employment Status:
- Employed full-time
- Employed part-time
- Unemployed
- Consultant
- Student
- Retired

(7) Political Identification:
- Far Left
- Left
- Center Left
- Independent
- Center Right
- Right
- Far Right
- Prefer not to answer

(8) Rank how strongly you agree with each of the following statements:
- Climate change is an important issue right now.
  —Strongly disagree
  —Somewhat disagree

—Neutral
—Somewhat agree
—Strongly agree
- Everyone should get their children vaccinated.
  —Strongly disagree
  —Somewhat disagree
  —Neutral
  —Somewhat agree
  —Strongly agree
- The federal government should be uninvolved in state affairs.
  —Strongly disagree
  —Somewhat disagree
  —Neutral
  —Somewhat agree
  —Strongly agree
- Donald Trump is a good president.
  —Strongly disagree
  —Somewhat disagree
  —Neutral
  —Somewhat agree
  —Strongly agree

B POST READING QUESTIONNAIRE

(1) Provide a one-sentence summary of the story you just read.
(2) How credible do you think the story was?
- Very uncredible
- Somewhat uncredible
- Neutral
- Somewhat credible
- Very credible

(3) Was this story written by a human or by a computer software?
- Human
- Computer software

(4) How engaging was this story?
- Very unengaging
- Somewhat unengaging
- Neutral
- Somewhat engaging
- Very engaging

(5) What is the political tone of this story?
- Far Left
- Left
- Center Left
- Independent
- Center Right
- Right
- Far Right

REFERENCES

[1] Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2109–2116. DOI: https://doi.org/10.18653/v1/N19-1216

[2] Alberto Barrón-Cedeño, Giovanni Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. *In Proceedings of the AAAI Conference on Artificial Intelligence* 33 (07 2019), 9847–9848. DOI: https://doi.org/10.1609/aaai.v33i01.33019847

[3] Charlie Beckett. 2019. *New Powers, New Responsibilities: A Global Survey of Journalism and Artificial Intelligence.* Technical Report. POLIS at the London School of Economics and Political Science. https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities/.

[4] Taina Bucher. 2017. "Machines don't have instincts": Articulating the computational in journalism. *New Media & Society* 19, 6 (2017), 918–933. DOI: https://doi.org/10.1177/1461444815624182 arXiv:https://doi.org/10.1177/1461444815624182

[5] William P. Cassidy. 2007. Online news credibility: An examination of the perceptions of newspaper journalists. *Journal of Computer-Mediated Communication* 12, 2 (01 2007), 478–498. DOI: https://doi.org/10.1111/j.1083-6101.2007.00334.x arXiv:http://oup.prod.sis.lan/jcmc/article-pdf/12/2/478/22316696/jjcmcom0478.pdf.

[6] Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL'96)*. Association for Computational Linguistics, Stroudsburg, PA, 310–318. DOI: https://doi.org/10.3115/981863.981904

[7] Junho H. Choi, James H. Watt, and Michael Lynch. 2006. Perceptions of news credibility about the war in Iraq: Why war opponents perceived the Internet as the most credible medium. *Journal of Computer-Mediated Communication* 12, 1 (10 2006), 209–229. DOI: https://doi.org/10.1111/j.1083-6101.2006.00322.x arXiv:http://oup.prod.sis.lan/jcmc/article-pdf/12/1/209/22316792/jjcmcom0209.pdf.

[8] Christer Clerwall. 2014. Enter the robot journalist. *Journalism Practice* 8, 5 (2014), 519–531. DOI: https://doi.org/10.1080/17512786.2014.883116

[9] Mark Coddington. 2014. Clarifying journalism's quantitative turn. *Digital Journalism* 3 (11 2014), 331–348. DOI: https://doi.org/10.1080/21670811.2014.976400

[10] Robert Edelberg and Neil R. Burch. 1962. Skin resistance and galvanic skin response: Influence of surface variables, and methodological implications. *JAMA Psychiatry* 7, 3 (09 1962), 163–169. DOI: https://doi.org/10.1001/archpsyc.1962.01720030009002 arXiv:https://jamanetwork.com/journals/jamapsychiatry/articlepdf/488197/archpsyc_7_3_002.pdf.

[11] Bryan Farnsworth. 2018. GSR and Emotions: What Our Skin Can Tell Us About How We Feel. iMotions. https://imotions.com/blog/gsr-emotions/.

[12] Edward A. Feigenbaum. 2003. Some challenges and grand challenges for computational intelligence. *J. ACM* 50, 1 (Jan. 2003), 32–40. DOI: https://doi.org/10.1145/602382.602400

[13] Andreas Graefe, Mario Haim, Bastian Haarmann, and Hans-Bernd Brosius. 2016. Reader's perception of automated computer-generated news: Credibility, expertise, and readability. *Journalism* 19 (02 2016). DOI: https://doi.org/10.1177/1464884916641269

[14] Jeremy Hsu. 2019. Microsoft's AI research draws controversy over possible disinformation use. In *IEEE Spectrum*. IEEE, New York, NY.

[15] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410* abs/1602.02410 (2016).

[16] Jaemin Jung, Haeyeop Song, Youngju Kim, Hyunsuk Im, and Sewook Oh. 2017. Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior* 71 (2017), 291–298.

[17] Spiro Kiousis. 2001. Public trust or mistrust? Perceptions of media credibility in the information age. *Mass Communication and Society* 4, 4 (2001), 381–403. DOI: https://doi.org/10.1207/S15327825MCS0404_4 arXiv:https://doi.org/10.1207/S15327825MCS0404_4

[18] Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, IEEE, Detroit, MI, 181–184.

[19] Jake Lever, Martin Krzywinski, and Naomi Altman. 2017. Principal component analysis. *Nature Methods* 14, 7 (2017), 641–642. DOI: https://doi.org/10.1038/nmeth.4346

[20] Bingjie Liu and Lewen Wei. 2018. Reading machine-written news: Effect of machine heuristic and novelty on hostile media perception. In *International Conference on Human-Computer Interaction*. Springer International Publishing AG, New York, 307–324. DOI: https://doi.org/10.1007/978-3-319-91238-7_26

[21] Bingjie Liu and Lewen Wei. 2019. Machine authorship in situ. *Digital Journalism* 7, 5 (2019), 635–657. DOI: https://doi.org/10.1080/21670811.2018.1510740 arXiv:https://doi.org/10.1080/21670811.2018.1510740

[22] Miriam J. Metzger, Andrew J. Flanagin, Keren Eyal, Daisy R. Lemus, and Robert M. Mccann. 2003. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association* 27, 1 (2003), 293–335. DOI: https://doi.org/10.1080/23808985.2003.11679029 arXiv:https://doi.org/10.1080/23808985.2003.11679029

[23] Clifford I. Nass, Matthew Lombard, Lisa Henriksen, and Jonathan Steuer. 1995. Anthropocentrism and computers. *Behaviour & Information Technology* 14, 4 (1995), 229–238. DOI: https://doi.org/10.1080/01449299508914636 arXiv:https://doi.org/10.1080/01449299508914636
[24] Graham Oppy and David Dowe. 2003. The Turing test. *Stanford Encyclopedia of Philosophy* 1, 1 (2003).
[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, Stroudsburg, PA, 311–318. DOI: https://doi.org/10.3115/1073083.1073135
[26] Shivam B. Parikh, Vikram Patil, and Pradeep K. Atrey. 2019. On the origin, proliferation and tone of fake news. In *Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, San Jose, CA, 135–140.
[27] Shivam B. Parikh, Vikram Patil, Ravi Makawana, and Pradeep K. Atrey. 2019. Towards impact scoring of fake news. In *Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, San Jose, CA, 529–533.
[28] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[29] Arnaud Rachez and Rodrigo Castro. 2017. Predicting political bias with Python. https://medium.com/linalgo/predict-political-bias-using-python-b8575eedef13 [Online; posted 18-October-2017].
[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog: Better language models and their implications. https://openai.com/blog/better-language-models.
[31] Monali Saraf, Tyrell Roberts, Raymond Ptucha, Christopher Homan, and Cecilia Ovesdotter Alm. 2019. Multimodal anticipated versus actual perceptual reactions. In *Adjunct of the 2019 International Conference on Multimodal Interaction (ICMI '19)*. ACM, New York, Article 2, 5 pages. DOI: https://doi.org/10.1145/3351529.3360663
[32] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. arxiv:cs.SI/1708.01967
[33] Jan-Philipp Stein and Peter Ohler. 2017. Venturing into the uncanny valley of mind—T influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition* 160 (03 2017), 43–50. DOI: https://doi.org/10.1016/j.cognition.2016.12.010
[34] Shyam Sundar. 2007. The MAIN model: A heuristic approach to understanding technology effects on credibility. In *Digital Media and Learning*. The MacArthur Foundation Digital Media and Learning Initiative, Chicago, 73–100.
[35] Andrew Thompson. 2017. All the news. https://www.kaggle.com/snapcrack/all-the-news.
[36] Alan M. Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing Test*. Springer, New York, 23–65.
[37] Van der Kaa, A. J. Hille, and Emiel J. Krahmer. 2014. Journalist versus news consumer: The perceived credibility of machine written news. In *Proceedings of the Computation+ Journalism Conference, Columbia University, New York*, Vol. 24.
[38] T. Franklin Waddell. 2019. Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly* 96, 1 (2019), 82–100. DOI: https://doi.org/10.1177/1077699018815891 arXiv:https://doi.org/10.1177/1077699018815891
[39] Ziang Xie. 2017. Neural text generation: A practical guide. *CoRR* abs/1711.09534 (2017). arxiv:1711.09534 http://arxiv.org/abs/1711.09534

Received December 2019; revised August 2020; accepted October 2020