Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking

An T. Nguyen[1], **Aditya Kharosekar**[1], **Saumyaa Krishnan**[1], **Siddhesh Krishnan**[1], **Elizabeth Tate**[1], **Byron C. Wallace**[2], **and Matthew Lease**[1]
[1]University of Texas at Austin [2]Northeastern University
atn@cs.utexas.edu byron@ccs.neu.edu ml@utexas.edu

ABSTRACT

Fact-checking, the task of assessing the veracity of claims, is an important, timely, and challenging problem. While many automated fact-checking systems have been recently proposed, the human side of the partnership has been largely neglected: how might people understand, interact with, and establish trust with an AI fact-checking system? Does such a system actually help people better assess the factuality of claims? In this paper, we present the design and evaluation of a mixed-initiative approach to fact-checking, blending human knowledge and experience with the efficiency and scalability of automated information retrieval and ML. In a user study in which participants used our system to aid their own assessment of claims, our results suggest that individuals tend to trust the system: participant accuracy assessing claims improved when exposed to correct model predictions. However, this trust perhaps goes too far: when the model was wrong, exposure to its predictions often degraded human accuracy. Participants given the option to interact with these incorrect predictions were often able improve their own performance. This suggests that transparent models are key to facilitating effective human interaction with fallible AI models.

Author Keywords

AI; Mixed-initiative; Fact-checking; Information Literacy

INTRODUCTION

In designating October 2009 as National *Information Literacy*[1] Awareness Month, former U.S. President Barack Obama drew national attention to a key 21st century information challenge:

> Though we may know how to find the information we need, we must also know how to evaluate it. Over the past decade, we have seen a crisis of authenticity emerge. We now live in a world where anyone can publish an opinion or perspective, whether true or not, and have that opinion amplified within the information marketplace.

Historically, we have relied upon information literacy education in our schools and libraries to teach our citizenry key critical reading skills, the importance of consulting multiple, independent sources, and to investigate the potential for underlying bias in whatever we read. For this reason, information literacy has been advocated as "a distinct skill set and a necessary key to one's social and economic well-being in an increasingly complex information society." [26]

However, the internet era presents new challenges to consumers of information. As information generation has accelerated, making sense of it has become increasingly difficult. While breaking down traditional barriers to authorship has democratized information exchange and dissemination, the massive growth of information production by less-established sources has created significant new challenges for readers in accurately interpreting and assessing the veracity of this barrage of content [47, 10]. The deluge of new online articles and sources means that is practically difficult for individuals to consistently manually cross-check sources. The rise of misinformation – unwitting or deliberate – has made it harder still for readers to tell fact from fiction.

In response, researchers in machine learning (ML) and natural language processing (NLP) have developed a variety of innovative new models that automatically fact-check claims [34, 31, 40]. However, these works have largely viewed fact-checking as a standard ML task in which the aim first and foremost is to achieve high predictive accuracy. While improving predictive accuracy is a laudable goal, we believe that consideration of the human element is equally important if such models are to be useful in practice.

We argue that fact-checking models must provide three key properties for practical use: (i) model transparency, (ii) support for integrating user knowledge; and (iii) quantification and communication of model uncertainty. Regarding (i), high predictive accuracy alone is insufficient; someone skeptical of online information is likely to be equally skeptical of any fact-checking tool. Indeed, many people distrust popular fact-checking services [4]. Consequently, a system must be transparent (and auditable) in how it arrived at its prediction so that a user can understand and trust the model. Concerning (ii), claim assessments will invariably rely at least partially on world-views (priors) pertaining to the perceived *a priori* credibility of claims and sources; a fact-checking system should enable explicit, individual specification of these, in turn providing a framework for users to easily inject their own views and knowledge into the system and realizing an integrated prediction that incorporates these. Finally, addressing (iii), system predictions ought to be presented as relative statements with respect to incomplete model specification and knowledge,

[1] https://en.wikipedia.org/wiki/Information_literacy

UIST '18, October 14–17, 2018, Berlin, Germany

ISBN 978-1-4503-5948-1/18/10…$15.00

DOI: https://doi.org/10.1145/3242587.3242666

rather than definitive judgments: an ML model should communicate its confidence in its predictions while accounting for potential sources of errors, empowering users to conduct their own in-depth inspection and reasoning.

In this work we present a mixed-initiative [19] approach that realizes the three properties above. We position automatic fact-checking as an assistive technology to augment human decision making. To intuitively complement and reinforce the information literacy skills that users bring to the partnership, we have designed the system to follow the same key steps advocated by information literacy education in order to estimate claim veracity: 1) find relevant articles (textual evidence); 2) assess each article's reliability and its relative support for the claim in question; and 3) assess claim validity based on this body of textual evidence. Moreover, by making the model's reasoning process transparent to the user and interactive, our interface has further potential to help teach and structure the user's own information literacy skills regarding the logical process to follow for assessing claim validity.

Our mixed-initiative approach to fact-checking blends human knowledge and experience with the efficiency and scalability of automated information retrieval and AI. Given a claim in natural language, the system first automatically finds and retrieves relevant articles from a variety of sources. It then infers the degree to which each article supports or refutes the claim, as well as the reputation of each source. Finally, the system aggregates this body of evidence to predict the veracity of the claim. Regarding user interaction, the (automatically inferred) source reputation and stance of each retrieved article can be changed via simple sliders to reflect user beliefs and/or to correct erroneousness model estimates. This, in turn, instantly updates the system's overall veracity prediction.

To evaluate our approach, we conduct three randomized experiments, designed to measure the participants' performance in predicting the veracity of given claims. The first compares users who perform the task with and without seeing ML predictions. Our results suggest that users tend to trust the model, even when it is wrong. The second compares a static interface to an interactive one in which users can fix or override model predictions. We found that users are generally able to do so, although this is less helpful when the model makes correct predictions. The last experiment compares a gamified task design to a non-gamified one, but we found no significant differences in performance and participation.

Contributions. We provide: (1) a novel mix-initiative approach to fact-checking, combining human and machine intelligence; and (2) a user study of our approach, revealing the practical promise and challenges of this human-AI partnership. To foster future work by others, we share our anonymized data, source code for significance testing, and an interactive demo[2].

RELATED WORK

Information Credibility

Someone skeptical of online information is likely to be equally skeptical of any fact-checking website or software. For example, many people are reported to distrust popular fact-checking services [4]. Just as we consider information credibility factors [16] in assessing a news article or website, therefore, we must also consider such factors in designing a website or web application to support fact-checking.

Some established best practices include: websites should be clearly organized and navigable, professional looking, well written, updated, and functional [16, 18]. More specific to this domain, sites and tools should be explicit about any potential sources of bias, e.g., by providing an "about" page to provide context, indicating any paid sponsors, discussing or posting an ethical code, and admitting when a mistake has been made. In relation to asking users to perform a potentially complicated task (e.g., consulting various uncertain evidence to fact-check a questionable claim) which heightens user uncertainty, a clean and usable website is even more important to further reduce cognitive load [43]. We have sought to adhere to the above best practices in designing our prototype web application.

Human Fact-checking and Information Literacy

Websites such as *Snopes* and *PolitiFact* have become increasingly important in providing expert fact-checking of popular claims. However, the reliance on human labor, particularly experts, does not scale to lets users check arbitrary claims.

Crowdsourcing-based fact-checking sites, such as *TruthSetter*[3], now also provide more scalable, peer-based assessment. Our design is inspired in part by recent research on crowdsourced fact-checking [41, 42] which suggests that

> ... the best remedy for propaganda and misinformation intended to manipulate public opinion is helping readers engage in critical thinking and evidence-based reasoning... [which] can have benefits well beyond identifying specific instances of "fake news" - it can teach users the critical thinking skills needed to detect and evaluate misinformation and fake news ...

In that work [41, 42], users can post claims, and other users can then share related sources, stances, and claims. We similarly structure the claim evaluation process through evidence collection and assessment. Whereas they develop a social, volunteer crowdsourced solution, we propose mixed-initiative approach between a single user and a machine learning system. Future work might usefully further integrate information literacy education with fact-checking of claims.

Another intriguing approach to teaching information literacy turns the process on its head, engaging people in gamified generation of fake news stories. By learning how to write fake news, participants were subsequently found to be less likely to believe or be persuaded by actual fake news articles [37].

AI-based Fact-checking

Many recent studies have explored the potential of AI for automated fact-checking [50, 49, 13, 34, 32, 45, 24]. These studies have primarily focused on model variants and techniques that increase the predictive performance of models (e.g., accuracy in predicting veracity). Hybrid work combining ML with

[2] https://github.com/thanhan/uist18

[3] https://truthsetter.com

crowdsourcing [24, 32] has similarly focused on predictive accuracy, without considering information literacy educational objectives, or exposure consequences, for the crowd [41, 48]. While some models do generate explanations for their predictions [34, 32], it remains unclear how users might interpret and interact with these predictions. Our work in this area is distinguished by our human-centered approach: our mixed-initiative design emphasizes the human-AI partnership, and our evaluation measures how use of such predictive systems impacts human inference in assessing claim veracity.

Recently, [22] present a visual analytic system for users to detect social media accounts that distribute misinformation. This is complementary to our goal of detecting false claims. Furthermore, although their system is inspired by a prediction model prediction [49], the users can not directly interact with that model to correct and override when its prediction is wrong.

Designing Human-AI Interfaces

As AI has been embedded within an increasing number of human facing applications, there has been a concomitant growth in interest in designing interfaces for humans to interpret and interact with these systems and predictions [1, 8]. In response, researchers have proposed a variety of novel interfaces for interacting with machine learning models, but these tend to require significant expertise on behalf of users [2, 25]. Others have developed interaction techniques tailored to specific tasks, such as image segmentation [11], image search [15], text classification [28], topic models [44], code search [36], and others. The fact-checking task is especially challenging, as the the system needs to present convincing evidence for its predictions, assuming users are (understandably) skeptical.

Beyond making ML more interpretable for people, there is also increasing appreciation for the greater potential capabilities that may be possible with hybrid AI-human collaboration. Jordan and Mitchel [21] recently opined about the kind of strengths each side of such a partnership, noting the potential to harness ML's ability to extract subtle statistical patterns from large datasets in concert with human skills in pulling these into plausible narratives informed by diverse perspectives. Such partnerships are now materializing even in creative endeavors, such as AI-human co-design of fashion [23, 46], creative writing [6, 35], and music composition [14].

To build such effective partnerships, Horvitz's suggestions for mixed-initiative design [19] remain relevant today, e.g.:

> We can enhance the value of automation by giving agents the ability to gracefully degrade the precision of service to match current uncertainty... We should design agents with the assumption that users may often wish to complete or refine an analysis provided by an agent.

Our system openly conveys its fallibility by showing the confidence of each prediction to the user. By clearly communicating rather than hiding this uncertainty, the system discourages users having over-confidence in the model or making poor decisions based on such a misunderstanding. Instead, the system presents the evidence it has in favor of its disposition, and leaves it to the user to consider this evidence in the context of their own prior knowledge and experience. We further provide an interaction mechanism by which users can inject their own beliefs into the system to refine its predictions.

Without users having a mental model of how the system combines evidence to predict claim veracity (i.e., model transparency), such interaction would not be possible. The importance of model transparency has been similarly reported in other studies, such as Kulesza et al. [27]'s report of a case study in which users responded positively to greater model intelligibility; as users learned more about the system through interaction, they became more satisfied with system output. As such, the study demonstrated that users valued going beyond "black box" ML and were willing and able to learn more about a system in order to use it more effectively.

Another area in which human-AI partnerships are being explored is interactive machine learning (IML) [1]. For example, recent work has investigated design and evaluation of IML systems with non-expert users [44]. Allowing users to alter system inputs and observe how outputs change in response is one technique for realizing model explainability [1]. Visual analytics [39] models similarly facilitate human decision making via model interaction. Following this principle, our model also offers fast incremental updates to predictions, enabling lay users to easily alter inputs via sliders and see immediate model updates of estimated claim veracity.

Unfortunately, adoption of best practices for usability in ML system design is not yet as widespread as one might hope [1]:

> ... machine-learning systems also often inherently violate many existing interface design principles. For example, research has shown that traditional interfaces that support understandability ... and actionability ... are generally more usable than interfaces that do not ... Many machine-learning systems violate both principles: they are inherently difficult for users to understand fully and they largely limit the control given to the end user.

This motivates greater collaboration between HCI and AI researchers, with both fields and their research products standing to benefit. HCI researchers have similarly endorsed the value for HCI to better understand and engage with AI [9, 17].

USER EXPERIENCE

This section describes the user experience we seek to cultivate with our mixed-initiative design. The main goal for our user interface is to realize transparency, in that users can understand how the system makes its predictions and thereby know when (and when not) to trust them. Below we discuss how we present our model outputs to users, including its final disposition regarding claim veracities and intermediate estimates of each article stance and source reliability.

User interactions with our system proceed as follows. The user first enters a claim (or selects an example claim, e.g., "Saudi Arabia has a new law that can force women to cover up their tempting eyes"). A list of articles relevant to this claim is then retrieved (along with the source of each article, i.e., the website where the article was published). Based on these articles, a prediction is made and presented to the user

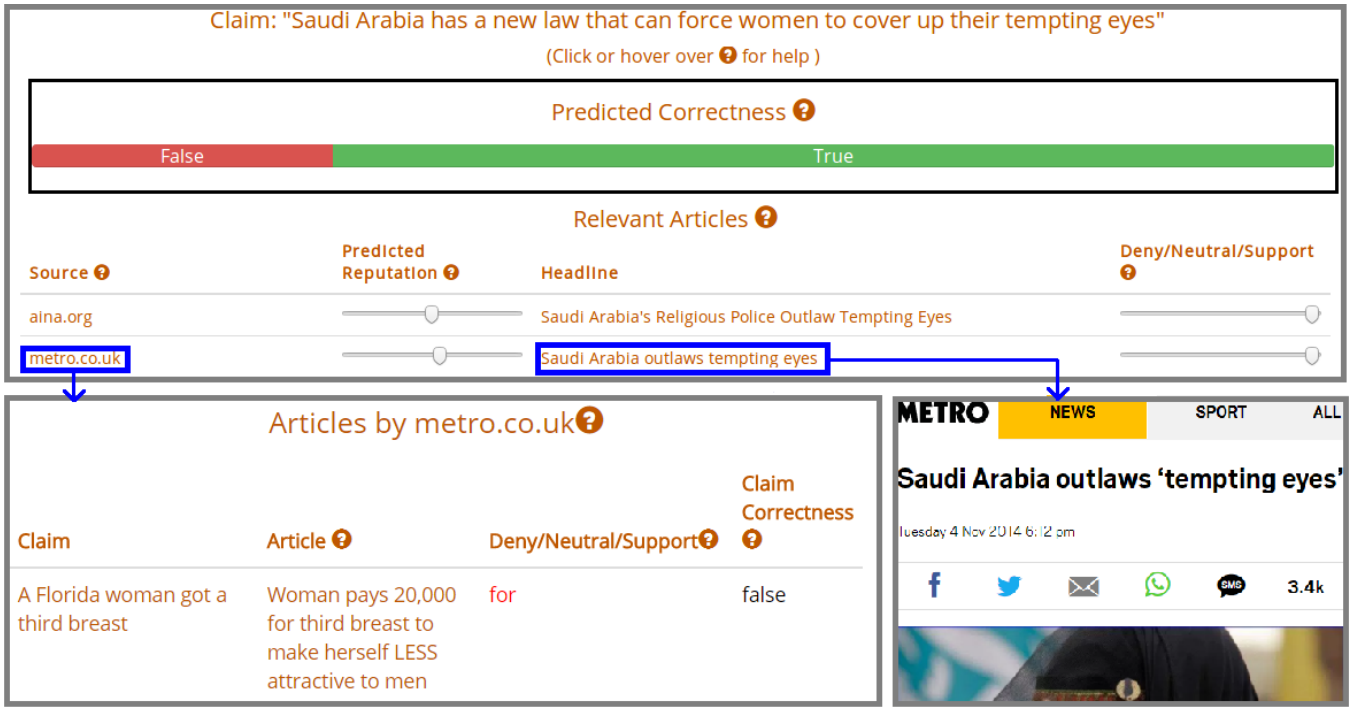


Figure 1: Top: the main results page, which includes the claim, its predicted correctness, and a table of relevant articles, their sources and inferred reputations and stances. Bottom left: sources link to pages that allow users to see the articles in our training set, which enables them to see why it has a particular predicted reputation. Bottom right: each headline links to the original article.

regarding the correctness of the claim; for example the model may be 80% confident that the claim is *true*.

Figure 1 shows a screenshot of the main results page for the example claim above. We display the claim and the system prediction regarding its correctness at the top. The interface emphasizes the textual evidence and reasoning underlying the overall claim estimate. In particular, we present a table of retrieved articles relevant to the claim, including the sources, headlines, and two predictions: the reputation of each source and the stance of each article headline. Each prediction is shown as a slider: reputation ranges between {*low*, *unknown*, *high*}, while stance ranges from {*deny*, *neutral*, *support*}.

A key feature of our interface is that users can change (override) the reputations and stances (inferred by the model) by moving the sliders. They can then observe how the prediction regarding overall claim veracity is affected. For example, for a headline associated with an article published by a reputable source, changing its stance from *support* to *deny* will tend to increase the chance that the claim is false. This interactive feature is beneficial in three ways. 1) It increases transparency: users can see how model predictions about each relevant article contribute toward the overall veracity prediction for the claim. 2) Users can access a more personalized prediction, for example by lowering the reputations of sources they believe are not credible. 3) They can correct the system's incorrect predictions, e.g., by changing predicted stances.

To further aid transparency, each headline is linked to the original article, allowing users to easily browse each article to see the textual evidence and assess the model's stance prediction. In addition, we generate a page for each source, and each occurrence of the source on a claim page is linked back to the source's generated page. When a user sees, for a given claim, that a relevant article comes from a particular source with a given prediction reputation, the user can consult the generated source page to learn more about the source and its predicted reputation. The source page lists all articles from that source in our training data, the claim associated with each article, the journalist-annotated article stance, and the journalist annotation for the claim's veracity. As before, users may click on each article to verify the stance label for themselves. By consulting the source page and seeing the various information about it, the user can thereby better understand the model's predicted reputation for the source: the more false claims supported by a source, the lower its reputation.

PREDICTION MODELS

In this section, we describe the predictive models underlying the above user experience. Our system's automated predictions are based on two machine learning classifiers: one for article stance, and one for claim veracity [34, 32]. Both classifiers are trained on the Emergent dataset [13], which contains 300 claims with 2595 relevant articles (on average there are 8.65 articles per claim). Each claim is labeled as *false*, *true*, or *unknown*. Article headlines are labeled by journalists as being either *supporting*, *neutral*, or *refuting* the claim.

Number	Claim	Veracity
1	Tiger Woods is serving a suspension from the PGA Tour after failing a drug test	False
2	There is a case of Ebola in Kansas City	False
3	The police officer leading the Charlie Hebdo investigation committed suicide	False
4	A 5-year-old boy was invoiced for missing a birthday party	True
5	The Indian government fired an employee who hadn't been to work in 24 years	True

Table 1: Five claims that we randomly selected for our user study from the test set of the Emergent dataset (excluding the attention check claim). Each claim is linked to the original Emergent's webpage showing the relevant articles and their stances.

The stance classifier accepts the claim and an article headline to predict the stance of the headline with respect to the claim. The veracity classifier operates over the outputs of the stance classifier for all of the relevant articles (and corresponding sources), yielding a prediction concerning the veracity of the claim. This veracity classifier explicitly models the reputation of each source. Sources that support true claims and deny false claims (in the training dataset) are given higher weights (i.e., more trusted). Both classifiers (stance and veracity) have an average accuracy of approximately 70%.

For inference and learning, we first train the stance classifier, then use its outputs to train the veracity classifier. For the stance classifier, we use the same text features as in [13], including bag-of-words (common n-grams), dependency parse, and paraphrase alignment (word embeddings are not used due to the reported negative impact). Each logistic regression classifer is implemented using Scikit-learn [33] with L1 regularization, Liblinear solver [12] and default parameters. While prior work has used joint training of the two classifiers in a graphical model framework [34, 32], we favor simplicity and speed to facilitate real-time user interaction.

The underlying model architecture of predicting article stances and using these predictions to estimate veracity is designed to improve transparency. While others have considered alternative architectures, for example deep neural networks [50], these can achieve good predictive performance but are less transparent, and so less amenable to interaction and supporting decision making. In contrast, we prioritize transparency over raw predictive performance. In particular, we rely on linear models in which individual terms have well-defined semantics, and we adopt a Bayesian view so that users may express subjective beliefs as priors imposed oer these variables. We operationalize this via a graphical user interface design.

USER STUDY

We conduct three experiments with participants from Amazon Mechanical Turk (MTurk). We required participants to have completed 1000 approved tasks with at least 95% approval rate. Participants were allowed to partake in only one of our experiments. The task takes less than 10 minutes on average and we paid $1.25, for roughly $7.50 per hour. While we did not collect participant demographics, because we post our task in small batches throughout the day, our participant demographics likely follow the general MTurk demographics reported in prior work [7, 38, 20]: mostly from the US or India, balanced gender, and younger than the working population.

In all experiments, participants predict the correctness of five randomly selected claims (**Table 1**) from the Emergent [13] test set. Note that our machine learning models have not seen any of these claims or the associated relevant articles.

An additional (sixth) claim served as an *attention check* (AC) [30]: a headline instructing participants to select a particular answer (e.g., "If you read this headline please select neutral"). While this AC was designed to filter out participants who did not pay attention to the task, we observed that many participants failing this check appeared to have honestly completed the task (e.g. many have accurate answers and helpful comments). We thus decided not to filter out any participants.

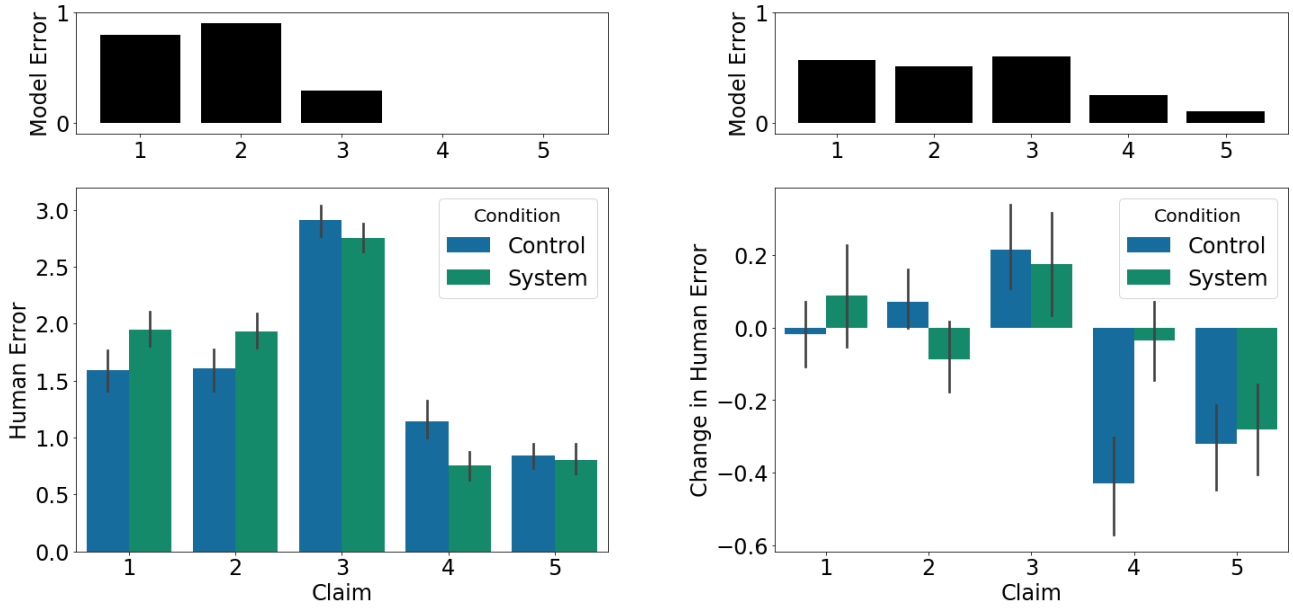
Experiment 1

This experiment tests whether system predictions help humans predict claim veracity more accurately.

Procedure: Participants are randomly assigned to one of two groups, *Control* and *System*. In both groups, users are first shown a screenshot of our results page (similar to Figure 1), but without the claim veracity prediction. In group Control, users are shown only the sources and headlines of relevant articles. In group System, they are also shown the source reputations and predicted article stances inferred by our system. The task for both groups is to evaluate the claim correctness, using a a Likert scale: *Definitely false*, *Probably false*, *Neutral*, *Probably true*, and *Definitely true*. After making this assessment, participants in both groups are shown the model's prediction concerning claim correctness and given the option to change their assessment. After completing this exercise for all claims, participants complete a short survey.

Results: We collected results for 113 participants (58 assigned to Control, 55 to System). We measure error by calculating the distance from the participants' responses to the correct answers. For example, for a (definitely) *false* claim, an assessment of *Probably false* corresponds to an error of 1, and *Definitely true* corresponds to an error of 4.

In **Figure 2**, we plot average errors over participants for each of the five claims, with the standard deviation displayed to characterize variance. On the left (**sub-figure a**), we show the error before the participants see the system's prediction. Firstly, we observe mixed differences across the two groups and five claims. Participants in group System were seen to show higher average prediction error on claims 1 and 2 and lower error on claim 4, while claims 3 and 5 show only small differences between the two groups.



(a) BEFORE seeing the system's correctness prediction.

(b) AFTER seeing the system's correctness prediction.

Figure 2: **Experiment 1**. *Control* group participants see a list of relevant articles with sources, while those in the *System* group see the same list along with predicted stances and reputations. Left: the errors of the *stance classifier* and of the participants before they see the overall veracity prediction. Right: the errors of the *claim veracity classifier* and the change in the participants' errors after being shown these predictions (a positive change means that human error increases).

Secondly, we observe variable accuracy of the automatic stance classifier. On claims 1 and 2, the stance classifier predicts the wrong answers for 80% and 90% of the relevant articles. Its error is 30% on claim 3 and 0% on claims 4 and 5. While average stance classifier accuracy is over 70%, this accuracy distributes unevenly across claims, with very low accuracy on two out of our five randomly selected claims.

For example, in inspecting the stance classifier on claim 1, we found a probable reason for its weak performance on this claim. This is an incorrect statement about Tiger Woods, a claim that is denied by many of the relevant articles. However, the stance classifier incorrectly predicts these articles' stances as being supportive. In the training set there are many articles about Tiger Woods that support a different (unrelated) claim. The classifier has incorrectly learned that the bi-gram 'Tiger Woods' indicates that the article is supportive. While a more sophisticated classifier might avoid this particular error, in general we should expect our AI systems to be imperfect.

Thirdly, we note the pattern of human error appears to roughly follow the system's stance classifier errors. It seems that when the stance classifier is wrong, participants are often misled by it, but when it is correct, it improves their predictions. This may suggest participants were overly trusting of our AI. While this is seemingly at odds with prior findings of users not trusting popular fact-checking services [4], it may stem from differences in participant demographics, participant incentives, or other factors of experimental design. For example, users in our study need to make predictions, instead of just saying whether they trust the fact-checking results. We return to this issue later in discussing study limitations.

In **Figure 2 (sub-figure b)**, we also plot the change in human error vs. claim after the participants saw our claim correctness prediction (a positive change means that human error increases). We observe that there are larger changes for errors made by participants in group Control: those who have not seen our stance predictions change their answers more than those who have. These response changes increase the error for some claims (e.g. claim 3), and decrease it for others (e.g. claim 4). This roughly corresponds to the errors by the veracity classifier, showing again that system predictions can both help users (when correct) or lead users to errors that reflect model fallibility or biases implicit in training data.

To quantify our results, we fit two Generalized Linear Models (GLMs): one for the data before participants see our system's correctness prediction, one after. We modeled human error as an ordinal response predicted by claim and participant as random effects, and the number of correct/wrong stance predictions by our system as fixed effects. Specifically, we use the clmm function of the R package ordinal [5] with the formula:

$$\text{Human.Error} \sim 1 + \text{CSP} + \text{WSP} + (1|\text{Claim}) + (1|\text{Participant})$$

where 'CSP' is the number of correct stance predictions that the human participant *sees*. It is 0 for all group Control participants who did not see any stance predictions. For group System, it is equal to the number of correct stance predictions for the claim. For example in claim 1, where the stance classifier is correct for 2 out of 10 articles, 'CSP' is 2 and 'WSP'

is 8 (WSP is similarly defined as the number of wrong stance predictions the participant sees). Also in the formula, the notation '(1|Claim)' means that Claim is a random effect and an intercept is estimated for each claim. For the data before seeing the correctness predictions, the results for the fixed effects are:

Coefficient	Estimate	SE	p-value (two-tailed)
CSP	-0.053	0.029	0.064
WSP	0.076	0.031	0.014

these suggest that seeing correct stance predictions (CSP) decreases human error while seeing wrong predictions (WSP) increases human error by a larger amount. Although the p-value for CSP is slightly larger than the 0.05 significant level, we consider that a solid evidence (the p-value is two-tail and includes the unlikely possibility that seeing correct stance predictions increases human error). After seeing the correctness prediction, the results are:

Coefficient	Estimate	SE	p-value (two-tailed)
CSP	-0.016	0.029	0.523
WSP	0.063	0.031	0.040

We can observe that seeing correct stances is now not as helpful because the participants can see the correctness prediction: in claim 4 and 5, many participants are able to lower their errors (Figure 2b). But seeing wrong stances is still harmful because these wrong stances cause the correctness classifier to make predictions with high errors (claim 1 and 2).

Experiment 2

This experiments assesses whether participants are able to interact with the system to inject their own knowledge, fix model predictions, and improve their own predictions.

Procedure: Whereas participants in Experiment 1 were shown only static screenshots, participants in Experiment 2 use our interactive interface. Participants were randomly assigned to two groups, *Control* and *Slider*. In the interface for group Control, all predictions (reputations, stances, and claim correctness) are fixed. However, those in the latter group (Slider) could change the (initially inferred) reputations and predicted stances, using the sliders, and observe how the prediction regarding overall claim correctness changes in response.

To encourage more attentive responses, we designed the task in this experiment as a simple game in which participants predict the correctness of a given claim (*false* or *true*) and indicate their confidence (0%, 5%, ..., 100%) in this response. Participants win points for correct answers, and lose points for incorrect answers. The number of points won or lost is proportional to their stated confidence. Participants may win up to 20 points on a given question; these are scaled linearly with the given confidence. For instance, a correct answer associated with a 75% confidence wins 20x75%=15 points (**Figure 3**). After a participant submits their prediction for a claim, we reveal the correct answer and the number of points that he or she has won or lost (for example "the correct answer is *false*, you have won 15 points"). The participant can then move to the next claim. After finishing the task for our selected claims, the participants have the option to continue working on other claims in the datasets, but this is not required. We hypothesized that participants who could see the consequence of their predictions in winning or losing game points may be more engaged and therefore make better predictions.

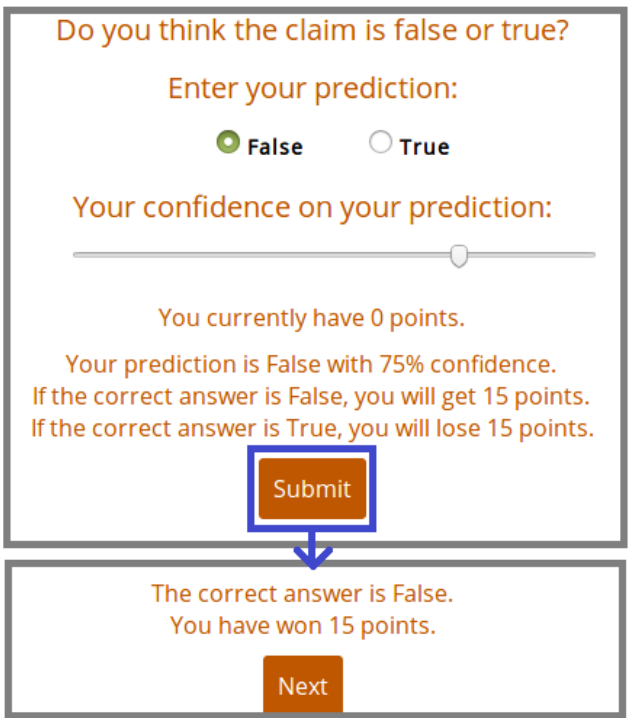


Figure 3: Top: In our gamified task interface, participants enter their prediction on the correctness of the claim and their confidence on the prediction. They win or lose points based on whether their prediction matches the correct answer. The number of points is 20 x confidence (20x75%=15). Bottom: after submitting, participants will see the correct answer.

Results: We collected results for 109 participants (51 assigned to Control, 58 to Slider). In **Figure 4**, we show the distribution of the participants' points as a boxplot, for each claim and condition. For claims 1-3, where the system is less accurate, participants in the Slider group (i.e., those able to use the slider-change feature) earn more points on average (assess veracity more accurately and/or confidently). For claims 4 and 5, where the system prediction is already accurate, boxplots reveal the Slider group participants have lower first quartiles, although the medians are still roughly the same. This suggests that some group-Slider participants are negatively impacted by the slider interface.

Within the 58 group-Slider participants, the sliders are used by 79% of the participants on claim 1, 59% on claim 2, and roughly 45% on claims 3, 4, 5. This decrease in use may be due to the variable error of the automatic stance classifier, or may reflect a familiarization effect as participants become accustomed to using the sliders.

We find no evidence of any difference in points between those who used the sliders and those who did not (p-value > 0.5

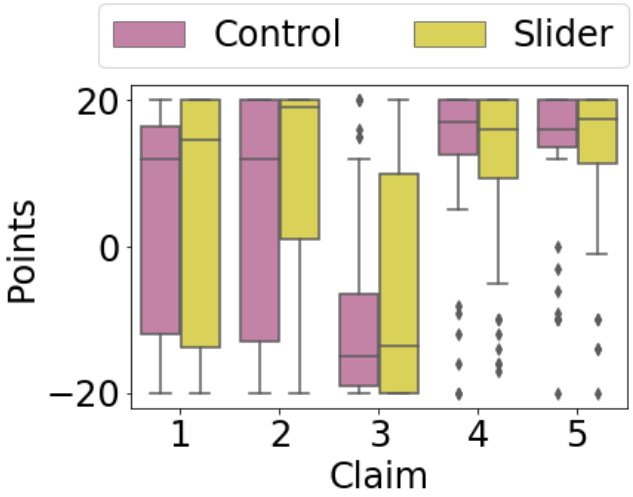


Figure 4: **Experiment 2**: boxplots of points earned for each claim in the two groups. In group Control participants cannot change the predicted stances and reputations; in group Slider they can. The diamonds indicate detected outliers (more than 1.5 Inter-Quartile-Range from the first or third quartile).

for all tests). This seems contradictory for claim 2, where we observe that group-Slider performs better than group-Control by a reasonable amount. If there is no difference between those who use the sliders and those did not, why was group Slider better for that claim? One possible explanation is that in claim 1, the Slider-group participants have learned to question the model (and not to blindly trust its predictions) so that they made better predictions on claim 2 (compared to group-Control), even for those who did not use sliders on claim 2. Although we have informed the participants (in both groups Control and Slider) in our task instruction that the model is only 70% correct, being able to experience the uncertainty in the model's predictions may help participants see the true quality of the model (by using the sliders and seeing how the predictions change).

Our GLM model for this experiment simply use the group assignment as the fixed effect:

$$\text{Points} \sim 1 + \text{Group} + (1|\text{Claim}) + (1|\text{Participant})$$

where 'Points' is treated as a continuous response. Using the function lmer in the package lme4 [3] and the tests in the package lmerTest [29], we have:

Coefficient	Estimate	Std. Err	2-tail p-value
Group Slider	1.721	1.232	0.166

We also fit the same model after removing the outliers (the diamonds in **Figure 4**) and find roughly the same estimate, with the p-value drops slightly to 0.122. The results suggests that participants in group Slider have higher points, although this is not significant at the 0.05 level. We believe that this relatively weak statistical signal is due to several reasons. First, there are large variances in human performance across claims and participants. Second, as discussed above, the sliders are not helpful (and sometimes even harmful) when the model is already correct. Third, the sliders were not used by as many participants as we expected. Finally, although the sliders are intended to be intuitive, they may require familiarization and a learning curve, and ultimately be found clearer to some users than others.

Experiment 3

This experiment investigates the degree to which 'gamified' aspects of Experiment 2 impact participant performance.

Procedure: We compare performance of two groups (to which participants were again randomly assigned). Both groups assess claim correctness and indicate confidence in this assessment. In the first group (Control) participants perform the task without the game (and so are not shown a number of 'points' they currently have or will get). In the second group (Game) participants complete the 'gamified' variant of the task.

Results: Across 106 participants (56 assigned to condition E, 50 to F), we find no significant differences between two groups in the number of points or number of extra tasks done (p-value > 0.5 for both). Consequently, we find that the game design did not impact participant performance (wrt. predictive performance or engagement) as hypothesized.

Survey Responses

In our post-task surveys across experiments, participants were generally positive about our tool, for example:

> "I thought it was really cool! I'd enjoy playing with this more if it wasn't during my work time."

Some participants expressed concern about the amount of information, the slider-change feature, and the correctness 'true' judgment (by the Emergent journalists).

> "there's a lot of data in this hit but not enough money to make it worth exploring"

> "Do not give me the option to tweak the deny, neutral, support rating as it led to some confusion regarding the task, however I was able to understand it once very quickly with practice."

> "It was very hard to understand. It seems on one task, I was 100% sure it was true and I was told it was false, I even followed links to verify the sources."

This first participant appears to be overwhelmed with the complexity of the fact-checking work, which is understandable. The second is not receptive to our slider-change feature, but also acknowledges that it became useful after some practice. The last participant seems to refer to claim 3, which no articles deny, but which the Emergent's journalists deemed false due to the reported event being exaggerated. This shows that fact-checking can be very difficult, and that many subtleties are lost by dichotomizing claims as either *true* or *false*.

DISCUSSION

We have presented a prototype system that enables users to interact with ML predictions for the challenging task of fact-checking (assessing claim validity). We designed the interface

so that users can know where the predictions are coming from. In some variants we allowed users to optionally override these predictions with their own beliefs or inferences. While our findings show that this human-AI interaction can be effective, they also suggest that caution should be exercised. Fallible ML models may make seriously wrong predictions (due to spurious correlations gleaned from potentially imperfect training data) that can in turn mislead users in some cases.

Limitations. This study presents what we believe to be intriguing results, but we note several important limitations. Firstly, our results in Experiment 2 are not significant at the 0.05 level, as we discuss, and should be interpreted with caution. Secondly, we have relied on MTurk participants, and different participant demographics or incentives may influence findings. For example, international MTurk workers may not be most representative of American news consumers or the most familiar and interested in American news. Thirdly, alternative plausible interpretations of the results exist. For instance, since MTurk workers are paid per task (rather than hourly), some workers may echo model predictions not due to trust but rather simple expediency of work. We acknowledge that possibility, though inspection of the data suggests such behavior only forms a small minority of what we observe.

It is also important to recognize that our work has the potential for *negative* impact as well. When the model makes errors, people may not recognize them, or could be even more confused by the introduction of AI modeling into an already confusing landscape of questionable sources and facts. While sliders support human-AI joint reasoning and allow users to correct modeling errors, they also create a new opportunity for self-constructed echo chambers, where correct model outputs can be manipulated at the whim of user bias. Were such "corrections" shared, one can easily imagine an adversarial setting where groups with competing interests seek to manipulate fact checking tools alongside their existing processes. Future work could consider designs to help users further recognize system limitations, interpret predictions with more caution, and explore the limits of their own knowledge and biases. For example, an interface could state the model's assumptions more clearly and ask users to confirm their understanding before they can see the predictions. Exploration of adversarial settings is also important to enable effective collaboration.

Our experiments include only claims for which we have elected to trust a "reference" veracity designation (by the Emergent journalists), and each has here been associated with a fixed set of relevant articles. As highlighted by the recent growth of work on algorithmic bias, our system learns only from the data it is given, with its accuracy and bias ultimately determined by that of the underlying data. An interesting direction for future work is to design for users to check their own claims, using relevant articles they find, and interact with other users' predictions.

In light of deeply divided views in political discourse and an increasingly ill-defined notion of 'truth', an assistive tool such as the one we present offers intriguing potential for brokering a more rational process in formulating individual beliefs and structuring debate among disagreeing parties [41]. If we can agree on the basic information literacy process to follow in assessing claims, and if we provide a structured process by which differing viewpoints can be precisely articulated and injected as prior knowledge into an automated system's reasoning process, perhaps we can create a space for more reasoned discourse. Instead of simply debating claims, two people with disagreeing views might sit down together and employ such a tool as a technological mediator. By alternatively injecting one another's viewpoints and beliefs as prior knowledge into the tool's reasoning process, we might come to more clearly understand the key evidence on which our beliefs disagree, and in so doing, gain additional insights into both our own beliefs as well of those who disagree with us.

ACKNOWLEDGMENTS

We thank Brent Biglin, Shravan Ravi, ChiaHui Liu, and Jyothi Vinjumur for their valuable contributions. The anonymous reviewers also provided terrific feedback that greatly improved this work. Last but not least, we thank the many talented Mechanical Turk workers for their participation in our study.

REFERENCES

1. Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
2. Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 337–346.
3. Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, and others. 2014. lme4: Linear mixed-effects models using Eigen and S4. *R package version* 1, 7 (2014), 1–23.
4. Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Trust and Distrust in Online Fact-checking Services. *Commun. ACM* 60, 9 (Aug. 2017), 65–71.
5. Rune Haubo Bojesen Christensen. 2018. R Package "ordinal". (2018). `https://cran.r-project.org/web/packages/ordinal/`.
6. Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces*. ACM, 329–340.
7. Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 135–143.
8. Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* (2017). `https://arxiv.org/abs/1702.08608`

9. Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 278–288.
10. Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web*. ACM, 341–350.
11. Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.
12. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research* 9, Aug (2008), 1871–1874.
13. William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *North American Chapter of the Association for Computational Linguistics*. ACL.
14. Rebecca Anne Fiebrink. 2011. *Real-time Human Interaction with Supervised Learning Algorithms for Music Composition and Performance*. Ph.D. Dissertation. Princeton, NJ, USA. Advisor(s) Cook, Perry R. AAI3445567.
15. James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*. ACM, 29–38.
16. BJ Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, and others. 2001. What makes Web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 61–68.
17. Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frederic Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, and others. 2016. Human-centered machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 3558–3565.
18. Aurora Harley. 2016. Trustworthiness in Web Design: 4 Credibility Factors. *Utg. av Nielsen Norman group. url: https://www. nngroup. com/articles/trustworthy-design* (2016).
19. Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.
20. Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk. (2010). NYU Working Paper No. CEDER-10-01. `https://ssrn.com/abstract=1585030`.
21. Michael I Jordan and Tom M Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260.
22. Alireza Karduni, Ryan Wesslen, Sashank Santhanam, Isaac Cho, Svitlana Volkova, Dustin Arendt, Samira Shaikh, and Wenwen Dou. 2018. Can You Verifi This? Studying Uncertainty and Decision-Making About Misinformation using Visual Analytics. In *12th International AAAI Conference on Web and Social Media (ICWSM 2018)*.
23. Natsumi Kato, Hiroyuki Osone, Daitetsu Sato, Naoya Muramatsu, and Yoichi Ochiai. 2018. DeepWear: A Case Study of Collaborative Design Between Human and Artificial Intelligence. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '18)*. ACM, New York, NY, USA, 529–536. `DOI: http://dx.doi.org/10.1145/3173225.3173302`
24. J. Kim, B. Tabibian, A. Oh, B. Schoelkopf, and M. Gomez-Rodriguez. 2018. Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. In *WSDM '18: Proceedings of the 11th ACM International Conference on Web Search and Data Mining*.
25. Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5686–5697.
26. Carol Collier Kuhlthau. 1987. *Information Skills for an Information Society: A Review of Research. An ERIC Information Analysis Product.* ERIC.
27. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10.
28. Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 1 (2011), 2.
29. Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 82, 13 (2017).
30. Catherine C Marshall and Frank M Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 234–243.
31. Ndapandula Nakashole and Tom M Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates.. In *ACL (1)*. 1009–1019.

32. An T. Nguyen, Aditya Kharosekar, Matthew Lease, and Byron C. Wallace. 2018. An Interpretable Joint Graphical Model for Fact-Checking from Crowds. In *AAAI*.
33. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
34. Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th Intl. Conference on World Wide Web*. 1003–1012.
35. Melissa Roemmele and Andrew S Gordon. 2018. Automated Assistance for Creative Writing with an RNN Language Model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. ACM, 21.
36. Xin Rong, Shiyan Yan, Stephen Oney, Mira Dontcheva, and Eytan Adar. 2016. CodeMend: Assisting Interactive Programming with Bimodal Embedding. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 247–258.
37. Jon Roozenbeek and Sander van der Linden. 2018. The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research* (2018), 1–11.
38. Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.
39. Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. 2016. Human-centered machine learning through interactive visualization. European Symposium on Artificial Neural Networks (ESANN).
40. Mehdi Samadi, Partha Talukdar, Manuela Veloso, and Manuel Blum. 2016. ClaimEval: Integrated and Flexible Framework for Claim Evaluation Using Credibility of Sources. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 222–228.
41. Ricky J Sethi. 2017a. Crowdsourcing the verification of fake news and alternative facts. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 315–316.
42. Ricky J. Sethi. 2017b. How citizen investigators can collaborate on crowdsourced fact-checking. (2017). The Conversation, November 5th. `https://theconversation.com/how-citizen-investigators--can-collaborate-on-crowdsourced-fact-checking-76890`.
43. Artemis Skarlatidou, Muki Haklay, and Tao Cheng. 2011. Trust in Web GIS: the role of the trustee attributes in the design of trustworthy Web GIS applications. *Intl. Journal of Geographical Info. Science* 25, 12 (2011), 1913–1930.
44. Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *23rd International Conference on Intelligent User Interfaces*. ACM, 293–304.
45. Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. In *Companion Proceedings of The Web Conference*. 517–524. Alternate Track on Journalism, Misinformation, and Fact-checking.
46. Kristen Vaccaro, Sunaya Shivakumar, Ziqiao Ding, Karrie Karahalios, and Ranjitha Kumar. 2016. The Elements of Fashion Style. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. 777–785.
47. Joyce Valenza. 2010. Web 2.0 meets information fluency: Evaluating blogs. `http://21cif.com/rkitp/assessment/v1n5/valenza1.5_blogeval.html`. (2010).
48. Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 275–284.
49. Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 647–653.
50. William Yang Wang. 2017. " Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 422–426.