Crowdsourced Detection of Emotionally Manipulative Language

Jordan S. Huffaker[1], **Jonathan K. Kummerfeld**[1], **Walter S. Lasecki**[1,2], **Mark S. Ackerman**[1,2]
Computer Science & Engineering[1], School of Information[2]
University of Michigan – Ann Arbor
{jhuffak,jkummerf,wlasecki,ackerm}@umich.edu

ABSTRACT
Detecting rhetoric that manipulates readers' emotions requires distinguishing intrinsically emotional content (IEC; e.g., a parent losing a child) from emotionally manipulative language (EML; e.g., using fear-inducing language to spread anti-vaccine propaganda). However, this remains an open classification challenge for both automatic and crowdsourcing approaches. Machine Learning approaches only work in narrow domains where labeled training data is available, and non-expert annotators tend to conflate IEC with EML. We introduce an approach, *anchor comparison*, that leverages workers' ability to identify and remove instances of EML in text to create a paraphrased "anchor text", which is then used as a comparison point to classify EML in the original content. We evaluate our approach with a dataset of news-style text snippets and show that precision and recall can be tuned for system builders' needs. Our contribution is a crowdsourcing approach that enables non-expert disentanglement of social references from content.

Author Keywords
Crowdsourcing; Media Manipulation; Rhetoric; Emotion

CCS Concepts
•**Human-centered computing** → **Collaborative and social computing;**

INTRODUCTION
Rhetoric that plays to people's emotions (e.g., fear-mongering rhetoric) can be an effective tool for inducing an emotional reaction in readers. Such reactions can cause "cognitive short-circuiting" [42, 44], resulting in the affected party taking actions or considering ideas they may otherwise disagree with (or even find repulsive). This is particularly true for emotions like fear and anger which increase susceptibility to tribalistic reasoning and inhibit empathy toward out-groups [31, 70]. A variety of actors exploit these effects for advertising [30, 52], increasing political influence [19], and amplifying misinformation, hate speech, and other harmful content [49]. We explore approaches for flagging potentially-manipulative emotional language in text in order to facilitate future counter-measures such as de-ranking or nudging information seekers away from hate speech and misinformation.

Specifically, we explore crowdsourcing methods that can overcome the challenges inherent in separating *emotionally manipulative language* (EML) from *intrinsically emotional content* (IEC) — that is, separating dramatic presentation that is meant to stir emotion in the reader (EML) from content that may be emotional on its own (IEC). For example, IEC might include an account like Camila's: "Being an illegal immigrant means living in fear of deportation....My 19-year-old brother was deported when I was 17....It's been seven years now that I haven't seen him and don't know if I ever will."[1] On the contrary, EML induces emotion with language cues, such as in a Fox News segment where Tucker Carlson claimed congresswoman Ilhan Omar is a "*living fire alarm*. A warning to the *rest of us* that we better change our immigration system *immediately. Or else*."[2] Carlson used EML to play to people's xenophobia, reminding viewers that Omar immigrated from Somalia (an out-group), and signaling to them that "foreigners" are taking over the nation by pushing progressive policies.

We show that automated approaches and standard crowdsourcing approaches fail to adequately make this distinction. It would be preferable to solve this problem using automated approaches due to their low cost and scalability. However, while we expect such approaches to be capable of finding the simplest cases of EML, they are constrained by the labeled datasets available to them and they lack the ability to understand social references. For example, it is difficult to proactively include phrases like "*lispy queer*", "*living fire alarm*", and "*bad hombres*" in training data because of their creative nature, yet they carry salient cultural implications that can induce intense emotional reactions. While crowdsourcing approaches are more expensive, they are cheaper than hiring experts and can be leveraged at any point during the 24-7 cycle of the information ecosystem by recruiting workers from a

CHI '20 April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00
DOI: https://doi.org/10.1145/3313831.3376375

[1] https://web.archive.org/web/20181210121052/https://www.manrepeller.com/2018/01/immigration-stories.html
[2] https://archive.org/details/FOXNEWSW_20190710_000000_Tucker_Carlson_Tonight/start/3540/end/3600

workforce such as Amazon Mechanical Turk (AMT). However, we find that non-experts struggle to classify EML in text because they lack the ability to disentangle sources of emotion (IEC and EML). When asked to find EML, they tend to conflate IEC in their judgment (an error we call *conflation error*). Cognitive psychology offers a possible explanation for this result: people are inclined to substitute hard judgments with easier ones, such as substituting EML detection with their affective state [39]. There are currently no known crowdsourcing approaches for dealing with this problem. Approaches for mitigating bias in crowdsourcing settings come closest, but they can only offer general warnings that still leave workers with the challenge of disentangling sources of emotion.

Instead, we propose a novel crowdsourcing approach we call *anchor comparison* that neutralizes the overpowering influence of IEC by measuring EML through comparison. We leverage workers' ability to identify and remove instances of language used to induce emotion in text to create a paraphrased "anchor text", then use that anchor as a comparison point with which to classify the original content. Our approach prevents extraneous factors within content from influencing classification by constraining judgment to differences from an anchor, ensuring that only EML is reflected in measurement. More generally, our approach is the first that can disentangle social references from content[3].

Anchor comparison classifies text for whether it contains EML while giving systems builders the ability to tune precision and recall, a useful feature for accommodating different applications. In our motivating interaction, we envision a system that uses EML detection to identify potentially-manipulative content and warns the user. For this type of system, a false positive would represent content that is flagged for EML despite containing no EML and a false negative would represent content that has EML going unnoticed by the system. False positives of IEC (conflation error) would be particularly problematic in cases where controversial content is unfairly flagged or punished, such as a post by a social justice advocate describing allegations of workplace harassment. We evaluate our approach by testing its ability to classify short text snippets as containing EML. We create a small dataset of short text snippets adopted from news articles, then systematically modify each snippet to create a version with heavy EML and one with very little EML while maintaining the same information between versions. We balance our dataset to include some stories with IEC and some without and measure classification performance with standard metrics (i.e., precision and recall).

In this paper, we make the following contributions:

- We identify a class of problems that involve disentangling social references from content (e.g., EML and IEC). They are too challenging for non-expert human annotators, who have a tendency to conflate content with references (an error we call *conflation error*).
- We introduce an approach, *anchor comparison*, that transforms classification problems into a comparison task to mitigate conflation of content (e.g., IEC) and social references. We leverage workers' ability to identify and remove instances of language used to induce references in text to create a paraphrased "anchor text", then use that anchor as a comparison point to classify the original content.
- We build a system that leverages anchor comparison to distinguish between intrinsically emotional content and emotionally manipulative language.
- We evaluate our system on a small dataset of short text snippets adopted from news articles and demonstrate both the limitations of existing approaches (i.e., automatic and standard crowdsourcing approaches) and the feasibility of our approach.

[3]By "social references" we refer to language that invokes connotations by overlaying parts of social and cultural contexts, in an often implicit manner. For example, "dirty crime-ridden cities" invokes a racist set of beliefs for many Americans, and "caravans" can not only invoke the literal meaning but also the Middle East. Social references can elicit multiple meanings for the majority of the audience. Many related terms for multiple meanings in language use exist in various literatures (e.g., multivocality [63], polyvalent performance [81], and dog-whistling), but none of these terms are precisely the same.

BACKGROUND

This work was motivated by a large set of related research areas. In this section we synthesize the literature that guided our system design: 1) emotionally manipulative rhetoric and how it effects information processing and 2) the strategies media manipulators use to shape the information ecosystem.

Rhetoric, Emotion, and Reasoning

Emotion is an integral part of how people perceive, process, and leverage information [44]. The role of emotion as a persuasive tool has been examined by scholars dating back centuries, notably including Aristotle who argued that pathos (appealing to an audiences' emotions) is one of three pillars for effective rhetoric [2]. Despite being an effective tool, scholars agree that the use of emotion becomes problematic when it is used to overwhelm a reader's ability to think rationally [9, 48, 55]. Informally this is referred to as an emotional appeal, an argumentative fallacy that encourages poor reasoning [22]. While there are many ways one can manipulate a reader's emotions, in this paper we focus on the ones that use emotional words and phrases to do so. Content that manipulates reader's emotions by adding an opinionated slant or by carefully including emotionally laden facts we leave for future work.

Psychology of Emotion. Emotion impacts people's decision making by triggering their fast-processing cognitive system, one of two systems people use to process information [40]. Lerner et al. describe this process and introduce the Emotional-Imbued Choice model of decision making [44]. Notably this model establishes that emotions shape the content and depth of thought people exert. For example, high-certainty emotions (e.g., anger) can lead to increased reliance on source credibility, leading to decreased attention to argument quality and higher usage of stereotypes and heuristics [6, 7, 79].

Importantly, manipulating emotions can impact how people behave because they are tied to *cognitive appraisal*, specific dimensions of cognitive state (i.e., how much attention we pay to the decision, how certain a person feels their actions

will lead to a specific outcome, etc.) that lead to predictable decision outcomes [72]. For example, inducing fear and anger increases vulnerability to tribalistic reasoning and can lead information seekers to take impulsive actions [31, 70]. One relevant work found that varying the content of a news article in the wake of the 9/11 terrorist attacks from an anger-inducing framing (discussing how Arabs were allegedly celebrating the attacks) to a fear-inducing framing (discussing how more attacks are to come) led participants to perceive more or less risk in the world and prefer policies that were more or less harsh on potential violators [43]. These studies provide a useful framework for understanding why emotion affects decision making. In the next section, we will discuss how adversarial actors have abused this framework to influence their audiences.

Media Manipulation

The rapid growth of social media has led to extensive changes in the media ecosystem, leaving it vulnerable to manipulation by a variety of actors [49, 76]. These manipulators have learned to game platforms' recommendation algorithms such as those that determine news feeds [80] and videos [45]. Given that two thirds of Americans get their news from social media [53], most people are exposed to content planted by adversarial actors and must discern which ideas conveyed are reasonable and which are harmful. When people come across content, they view it through the lens of their social identity. Content that threatens, acknowledges, or confirms their identity can create an emotional reaction rather than a reasoned one [64], and it is this response that is most responsible for the spread of misinformation, hate speech, and harassment-inciting content [32, 50].

Actors have learned to maximize the virality of their content by playing to people's emotions. For example, Vosoughi et al. found that viral hoax tweets were more novel and evoked stronger feelings of surprise and disgust than non-hoax viral tweets [80]. Additionally, Song and Gruzd found that anti-vaccination content was much more popular and was more likely to be labeled under entertainment categories than pro-vaccination content [75].

Finally, actors often play to their audience by leveraging social appeals and cultural contexts. For example, Lewis found that a network of far-right YouTube channels give credence to one another by maintaining close social ties through hosting one another in their videos [45]. While many of these YouTubers hold contradicting beliefs, they mask their inconsistency by highlighting shared values, like their understanding of internet culture and their reactionary stance toward current events and ideas (e.g., feminism, social justice, and left-wing politics). Looking closer at many click-bait "fake news" stories, Marwick found that successful articles often connected to "deep stories", or the larger narratives readers often hold (e.g., that the rural states are neglected in favor of big cities) [50]. Now that we have described how actors use emotionally manipulative rhetoric to manipulate the information ecosystem, we will describe related work that attempt to mitigate these efforts.

RELATED WORK

In this section we will describe existing approaches for countering the effects of manipulative rhetoric and for setting up interventions at scale. We will first contextualize our approach within the body of intervention strategies, then we will describe specific approaches that might be leveraged to detect EML within text.

Mitigation Strategies

Prior work has explored two primary ways to mitigate manipulative rhetoric: 1) blocking or reducing the reach of explicitly-harmful content and 2) nudging people away from potentially-harmful content with warnings.

The first strategy has been widely adopted by social media platforms to limit the reach of blatant click-bait articles by removing them or by down-ranking them in search query results and recommendation algorithm suggestions [5, 10]. However, platforms have a variety of reasons to be hesitant to use this power including legal, financial, and political concerns [27]. Therefore, they typically reserve blocking and down-ranking for only the most extreme content [11].

The second approach, briefly adopted by Facebook [61], is to display a message to potential readers of an article that it is "disputed", "false", or something similar in order to dissuade people from believing it. In controlled lab settings, this approach has been shown to slightly reduce belief in the article [16]. However, people can become dependent on such labels after getting used to them, leading them to be more likely to accept un-flagged false articles as true than they would otherwise were there no flagging [63]. More effective interventions prime information seekers to think more carefully about the content they read [8] or warn them about specific strategies used to manipulate them [17].

Since these intervention approaches require manipulation efforts to be detected at scale to perform interventions for users of social media, many have sought to build automatic detectors. Detecting manipulative content can be done by identifying a variety of credibility indicators [87], including source [51] and content-based indicators [3, 58, 65, 71]. Often, credibility indicators are buried in context and require social and cultural knowledge to uncover them, making them challenging to identify. For these scenarios, prior work has demonstrated that collective behavior of information seekers can be a powerful proxy for uncovering credibility scores, including collective attention [57, 85]. This work demonstrates that combining time series and aggregate attention behaviors can be used to predict the credibility scores of tweets, a finding that might also extend to news article shares. While aggregate information seekers' behavior may be used to uncover other credibility signals, this strategy can only measure the effects content has on their information seeking, and would be unable to detect more nuanced signals that cause such behavior. Instead, we propose a proactive approach that involves looking for a particular credibility signal (EML) before the content has the chance to impact information seekers. Next we describe related work for carrying out this proactive approach within text at scale.

Possible Approaches

Next, we look to relevant work in natural language processing (NLP) and crowdsourcing to find a potential classifier for EML; however, we find gaps in both approaches that limit

their performance. In NLP, we find related work is constrained by available manually annotated data and methods are often dependent on key terms, lacking the ability to understand social references. Within crowdsourcing we find related work in making subjective judgments and bias mitigation, but we find a lack of work in crowdsourced disentanglement.

Natural Language Processing. While we are unaware of work in NLP on this specific problem, there has been extensive work on related classification tasks, such as sentiment analysis and emotion detection. Sentiment and emotion are a key part of rhetoric, and can be difficult to separate from the intrinsic content of text, similarly to the EML and IEC conflation we consider. However, most sentiment analysis research has focused on product or movie reviews [46, 74, 88]. Recent work has broadened the types of texts considered to include a range of topics on Twitter, but all labeling is completed by simply asking crowd workers to judge sentiment on a scale [18, 68]. Work on emotion detection has considered a wider range of text sources, but methods rely on expert annotated data that is collected with heuristic approaches to focus on examples expressing emotion [12].

Some of the most widely cited work in sentiment and emotion classification is the creation of lexicons specifying words with particular emotional content [23, 78]. While effective for the tasks mentioned above, these resources fall short for the manipulative rhetoric we consider here, in which words are used creatively, exploiting references to recent events, dog whistles and more. Prior work has explored bootstrapping existing context-based approaches by using knowledge bases or video footage in order to evaluate the meaning of these references [13, 77], but such solutions do not scale to provide knowledge of the history, culture, and community norms necessary to evaluate references that are grounded in social and cultural contexts —like many of the emotive phrases seen in news articles [26]. Systems that do not rely on such lexicons instead rely on large collections of manually annotated data. These are expensive to create with experts and crowdsourcing faces difficulties that we discuss in the next section.

Crowdsourcing. Crowdsourcing is a common approach for overcoming the limitations of automatic approaches, often applied to gather training data [38, 73] or to integrate human intelligence into computational processes [4, 66]. Again, there are significant gaps in the crowd's ability to disentangle social references. Traditional approaches to achieving quality responses are to improve task instructions [47, 82], train or screen workers [56], or to decompose the task into subtasks [41]. Newer approaches leverage the crowd's reasoning skills to improve results [14, 21]. While these approaches are useful for addressing challenges like task misunderstanding, low quality workers, and groupthink, they were not designed to help workers disentangle EML from IEC —which we show is still hard for even high-performing workers who understand the instructions clearly. Decomposing the task might make more sense, but there is no obvious way to decompose subjective judgment tasks such as classifying EML yet.

Alternatively, a line of work has explored the crowd's ability to make subjective judgments and mitigate bias. The crowd is particularly well suited for making subjective judgments because they can leverage social and cultural knowledge to predict how others might answer the same task [15]. Prior work has leveraged the crowd's ability to make these judgments to build emotion lexicons [60], assess image quality [67], and curate content based on personalized preferences [54], among other applications. Additionally, one might think of disentanglement as a bias mitigation problem. Prior work has explored bias mitigation strategies including those that warn workers of potential biases [34, 35, 69] and those that leverage statistical methods to debias results after-the-fact [37]. While this work helped us form an understanding of our problem, we note that disentanglement differs from bias mitigation in that specific biases cannot be known ahead of time. Without knowing biases, intervention-based approaches cannot be easily applied, as they would result in vague warnings to workers. We introduce an approach that adds to subjective judgment literature by enabling crowdsourced disengagement of social references such as disentangling EML and IEC.

PROBLEM

Before we describe potential classifiers to detect EML, we will define our problem more precisely, describe a small dataset we created to evaluate a variety of detection approaches, and explain how we measure performance. For the purposes of this paper, we treat EML detection as a classification problem and we specifically target extraneous language intended to induce an emotional reaction in the reader. We leave related tasks such as identifying specific EML words and phrases within a larger body of text and detecting biased reporting due to skewed facts for future work.

We envision two main applications that would benefit from the simple classifier as we have described: one that uses detection for intervention purposes (e.g., personalized nudges) and one that uses detection for content moderation purposes (e.g., de-ranking offending content or flagging content for site managers). We are particularly motivated by the first of these applications, as EML detection could be a useful backend for a system that performs inoculation interventions (e.g., [17]) or a system that points out specific EML words and phrases in the original content. A false negative would represent content that has EML remaining un-flagged, which in our scenarios, might result in lower user confidence for the overall system. On the other hand, a false positive would represent content that is incorrectly flagged for EML, which could lead to over prompting of the user or unfair punishment of content.

As a first step, we focus on classifying short text snippets (<200 words) to ensure that workers can read and comprehend the text in the timespan of a microtask. Future work will extend to longer text such as full news articles, posts, or threads. We explore two factors that may influence classification:

1. *Emotionally Manipulative Language*: Whether there is highly emotive language that adds no informational value to the text. This is the factor we seek to detect.
2. *Intrinsically Emotional Content*: Whether the information conveyed is emotional itself, regardless of the language used to convey it. Information seekers frequently must parse emotional stories from manipulative language. The

	¬IEC		IEC	
	¬EML	*EML*	*¬EML*	*EML*
# snippets	5	5	5	5

Table 1. Our test dataset is balanced across four conditions.

common news media trope "if it bleeds, it leads" hints at the prevalence of dramatic stories in media [84].

Dataset

The authors developed a dataset of twenty news-style text snippets adopted from 10 news articles[4]. For each news article, we created a shortened version that includes heavy EML and a version that includes very little EML, creating a pair. We created each pair so that both versions had the same information, with the only difference being EML. We picked five news articles that contained IEC and five that did not, making the dataset balanced with five snippets per condition (Table 1).

We evaluated the quality of our dataset by hiring a journalist and a member of the editorial staff for a nationally prominent news magazine to rate each text snippet on a 5-point Likert scale along three dimensions. First, we confirmed the main factor of our dataset by asking reviewers to rate "How much does the paragraph intentionally stir emotion in the reader?". Second, we confirmed our variation of IEC by asking raters to assess "How much emotion is intrinsic to the information conveyed in the paragraph?". Finally, we sought to confirm that some snippets are publishable in reputable news sources by asking raters to agree or disagree to: "If the facts were correct, this is something that would be publishable in a reputable news source."

The first author met with each rater separately and followed a three part procedure for each snippet in our dataset: 1) we asked the raters to rate the snippet on the three dimensions mentioned, 2) we asked the raters for the reasoning behind their answers to ensure that they understood the meaning of each question, 3) we allowed the raters to change their answer after discussion. We reached a high Fleiss' kappa score with the two raters and the first author about which of the snippets in each pair used more EML (k = 0.87), which snippets had a "decent amount"[5] and which had more intrinsic emotion (k = 0.60)[6], and which of the text snippets in each pair were more likely to be publishable by a reputable news source (k = 0.72).

Measures

Our main outcome measures are standard classification metrics (i.e., precision and recall), focusing specifically on conflation error (false positive rate of IEC). Additionally, we define two more metrics that we will use later in this paper to evaluate intermediate outcomes: *EML level* and *distortion*. EML level is coded on a 4-point scale and indicates the relative amount of EML in a text snippet with 4 being the maximum and 1 being the minimum. In each snippet pair, we define the snippet with EML to be level 4 and the one with no EML to be level 1. Additionally, we use the metric distortion to code whether a change to a text snippet has changed its informational content. We use this dataset and described measures to evaluate various classification approaches.

[4]Can be accessed online at: https://doi.org/10.7302/yhpy-e679

[5]This was the middle option on a 5-point Likert scale.

[6]One of the raters found some of the text snippets more personal (and thereby having more intrinsic emotion) than the other rater and first author found. To ensure these differences would not affect our results, we ran our evaluation with the codes provided by the differing rater and found only marginal difference in the results of our standard crowdsourcing approach.

BASELINE STUDY

In order to understand the limitations of existing approaches to classifying EML, we first explore five logical baselines and find that none achieve satisfactory performance. In particular, we explore four automated approaches and a simple crowdsourcing approach.

Baselines

At first glance, it may seem that automatic approaches (e.g., machine learning) can be trained to sufficiently classify EML due to their success in classifying text in similar domains (e.g., emotion detection and sentiment analysis). For this reason, we evaluate state-of-the-art performance for *sentiment analysis* and *emotion detection* by classifying snippets from our dataset.

Emotion detection is the task of classifying the intensity of typically 4-6 emotions induced by a piece of text. Existing approaches make use of an emotional lexicon to match words with the emotion they commonly evoke [59], then aggregate those emotions to determine the overall emotion evoked by the paragraph. Specifically we evaluate:

IBM Tone Analyzer [1]: This classifier was trained on Twitter customer service data and uses a support-vector machine (SVM) to classify emotion on a scale from 0-1 for anger, disgust, fear, joy, and sadness.

EMPATH [24]: EMPATH is a tool for coding lexical categories (such as emotion) in large-scale datasets (similar to LIWC). The tool "adapts" to new dataset domains by enabling users to seed additional lexical categories, and then by recruiting workers from AMT to find related terms in the dataset. We use EMPATH to code emotion words.

We also considered sentiment analysis, which is the task of classifying positive, negative, or neutral feelings expressed in text. While sentiment analysis is less related to EML detection than emotion detection, the datasets for sentiment analysis are better developed than those for emotion detection are, and sentiment analysis should still be capable of detecting large attitude-slants. We explore:

VADER [36]: VADER was designed to generalize to a variety of domains by combining features from a sentiment lexicon with five rules. Its rules were created from a qualitative coding of tweets and its simplicity makes the model perform well without large-scale training data, unlike most other approaches.

BERT with Fine-Tuning [20]: we trained a three-way classifier on top of the BERT-Base uncased pretrained model, using the 2017 SemEval sentiment analysis task data [68]. This approach takes into account the context of words by using a pretrained language model (BERT) achieving an F1

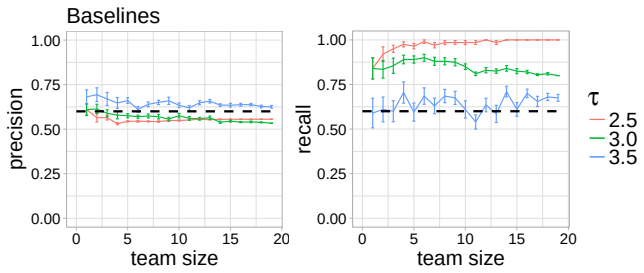

Figure 1. Precision and recall of the baselines. The dashed line indicates the best performance of our automatic baselines and "τ" indicates different decision boundaries on a 5-point Likert scale.

score of 0.636, slightly below the state-of-the-art for the dataset (F1 = 0.677)[7]. We used a Twitter dataset because it offered more diversity than the movie and customer review datasets, making it more likely to include information about current events and culture.

Finally, we consider a standard crowdsourcing approach for classifying EML that asked workers from Amazon Mechanical Turk (AMT) to rate text snippets on a 5-point Likert scale[8], takes the average of those ratings, then uses a decision boundary to determine positive or negative classification. For our evaluation, we recruited workers from AMT using Legion-Tools [29], presented them with a single text snippet from our corpus (requiring unique workers for every task), and asked them to answer the question "How much does the paragraph intentionally stir emotion in the reader?"[9].

Results

We find that the automated baselines only perform marginally better than a random baseline and the standard crowdsourcing approach performs slightly better than the automatic baselines for recall, but has similar precision (Figure 1). In particular, we find the IBM and BERT baselines have a high false positive rate while EMPATH and VADER had balanced error rates [10]. Additionally, we see that the crowdsourcing baseline had high conflation error (misclassifying IEC text snippets as having EML), explaining the low precision.

Automated Approaches. While it may be possible to improve the performance of machine learning approaches by training on a dataset explicitly labeled for EML, we argue that their performance will remain limited for three reasons: 1) generalizing knowledge to examples not explicitly trained for remains a challenge for even state-of-the-art models, 2) when these classifiers are deployed in the wild, adversarial actors will be highly motivated to find exploits, and 3) hiring experts to create labeled datasets is expensive and time consuming.

[7]We attempted to train the state-of-the-art model but experienced issues in the training process. Since our model performs similarly on the Twitter dataset, we believe that its results should be representative.

[8]Likert scale ratings are standard for related tasks [36].

[9]In a prestudy, we tried many ways to word this question in order to achieve better results. We found that wording only marginally effected results, and that our findings hold despite the choice.

[10]The IBM Tone Analyzer and EMPATH detectors output continuous scores. We convert these scores to classification categories based on whether they cross a decision boundary. For all decision boundary values, accuracy did not exceed 60% for either detector.

First, machine learning approaches are limited in their ability to generalize knowledge beyond what exists in training data, making them vulnerable to novel patterns and references not explicitly trained for. For example, phrases such as "*lispy queer*", "*living fire alarm*", and "*bad hombres*" are rare manipulative phrases that are unlikely to show up in training data, and so would likely go unnoticed. The use of cultural references are common, constantly changing, and can take many different forms, making it nearly impossible to include all of them in training data. As this limitation has become increasingly better known, scholars have started developing methods to test classifiers for their generalizability by developing "adversarial datasets", modified versions of existing datasets whose changes do not effect human performance but often tank machine performance [33, 62, 86]. We expect such datasets will be important for assessing future EML classifiers.

Second, prior work has found that adversarial actors are highly motivated to exploit algorithms that are deployed in the wild, often in a coordinated manner [28]. In particular, actors actively manipulate search engines to amplify their content by finding "data voids", search terms with limited data available to populate search results, then by posting extremist content that uses those search terms. For our envisioned interactions (i.e., content moderation and intervention systems), we expect that actors will also be motivated to find novel exploits, making it important that backend EML classifiers are reasonably robust against these attacks.

Finally, hiring experts to create a labeled dataset for EML detection is expensive and time consuming, making it infeasible to fix the previously described problems by creating massive annotated datasets. Since, as mentioned, cultural references are constantly in flux, over time new annotations would need to be gathered to keep up detection quality amid new current events and discourse themes. In addition to performing better, a crowdsourcing approach would be more cost effective than hiring experts, and thereby might be used to cheaply create a large-scale dataset for EML detection.

Standard Crowdsourcing. The simple crowdsourcing approach failed because workers tended to conflate IEC with EML. A chi-square test comparing worker ratings for snippets with IEC and no EML with their ratings for snippets with no IEC or EML confirmed that the tendency to conflate was significant $\chi^2(4) = 32.49$, $p < 0.001$ (effect size = 0.44).

This finding is supported by the theory of the affect heuristic in psychology, which contends that people commonly use their emotions as a cognitive shortcut to make judgments (i.e., how they are emotionally affected by the judgment affects judgment) [25]. While we cannot be certain that the affect heuristic is the only cause of conflation error in our context, it does offer an explanation for our results: workers are affected by IEC and are substituting the EML detection with their affective state by rating IEC highly. Unfortunately, this problem is not easily solved, as prior work has found that people struggle to disentangle sources of emotion even after being warned about the potential to attribute emotion to the wrong source [81].

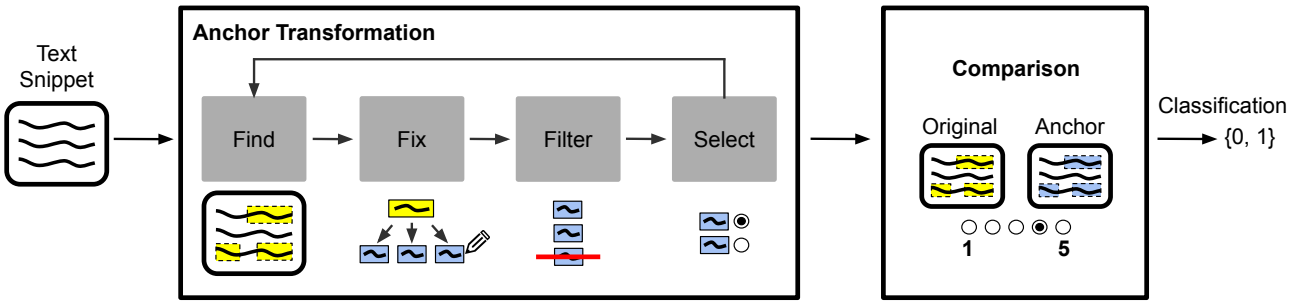

Figure 2. Our system splits the task of classifying EML into two parts: anchor transformation and comparison. Anchor transformation involves removing EML from the original text snippet to form an "anchor". This is done in four steps: finding portions of text that contain EML, suggesting possible edits to remove EML from each portion, filtering edits to remove those that introduce distortion, and selecting the best edit from a group.

Given that automatic approaches did not work sufficiently for this problem, and a standard crowdsourcing approach led to the challenge of disentangling sources of emotion, we developed a new crowdsourcing approach. In the next section, we will describe how we overcome the limitations of these baselines by transforming the judgment task into a comparison problem, thereby limiting the influence of IEC on final classification.

ANCHOR COMPARISON

In this section, we will describe a system we created that leverages an approach we call *anchor comparison* to mitigate the overpowering influence of IEC on worker judgment. Our approach decomposes the problem into two pieces (Figure 2): 1) anchor transformation which involves coordinating workers to remove specific instances of EML in the text to create an "anchor" version of the original and 2) comparison which involves measuring the difference in EML between the original and the anchor on a 5-point Likert scale. It thereby turns the classification problem into one of comparison, enabling a task that was previously considered to be atomic to be decomposed.

To explain why anchor comparison works, we will build off our previous psychological analysis of the problem. As we have previously explained, workers likely conflate IEC with EML because they use their affective state (how they feel) to decide how to classify the content [25]. IEC makes them feel strongly, which is then reflected in their Likert scale ratings. Our approach works by anchoring workers' affective state in the anchor text and measuring only the difference from the anchor to the original text. Differences from the anchor may still affect workers, but this affect would be due to EML since IEC is held constant. Our approach builds upon reference-based crowdsourcing approaches such as [83] in that we leverage the crowd to create the point of reference. In the next sections, we will walk through each component of our system. While we do so, we will describe a study we used to measure its performance and set parameters. We will conclude by describing two key tradeoffs our system affords.

Anchor Transformation

Anchor transformation is the process of transforming text into an "anchor", a paraphrased version of the original text that has been revised to remove EML. The problem shares many characteristics with copyediting with one key difference: particular sensitivity to "distortion" errors, where workers alter the information content in the text in an attempt to remove EML. Distortion can lead to false positives (and conflation error) in the comparison step by creating a fabricated difference between the original text and the anchor.

To enable explicit control of distortion, we build upon the current state-of-the-art for crowdsourced copyediting (Bernstein et al.'s Soylent [4]) by adding a "filter" and "select" step. The resulting system consists of four steps: finding EML words and phrases, suggesting potential edits to remove EML from those phrases, filtering out suggested edits that have distortion, and selecting the best edits from those remaining. For each step, we hire workers to complete the task in parallel (hiding other worker's responses to avoid potential groupthink) and pay them at an hourly rate of $10 USD/hour. We will describe each step in detail below.

Find. The find step involves identifying parts of the text that have EML by enabling workers to highlight portions of the text, then aggregating based on highlight overlap. Specifically, we ask them to "Highlight dramatic[11] words and phrases in the [text]." In our component study we found that a 20% worker overlap threshold is the optimal value for maximal overlap with our annotations of the text and that performance plateaus at approximately 5 workers at 75% word-wise agreement.

Examining the portions highlighted by workers, we notice four types of highlights. The first are highlighted portions like "tragic deaths" and "sweet and unsuspecting American children" that can be fixed by simply removing the unnecessary verbiage (e.g., "tragic" and "sweet and unsuspecting"). Secondly, some highlights were made of entirely IEC like "Panama fungal disease threatens future crops" and require no editing. Thirdly, some highlights contain both EML and IEC like "Gary was swept by a wave of grief". These highlights require clever rephrasing in order to maintain information while removing EML (e.g., rephrasing to simply "upset"). Finally, some highlights contain both EML and IEC, but are particularly challenging to rephrase in a way that maintains

[11] We switched to using the word "dramatic" instead of "emotional" after analyzing workers' qualitative explanations for their responses in a pre-study and noticing that many workers interpreted "emotional" in the instructions to mean "that I had an emotional reaction to" instead of our intended meaning "that the author was *trying* to get me to react emotionally to". We tried a variety of wordings and found that "dramatic" yielded the best results.

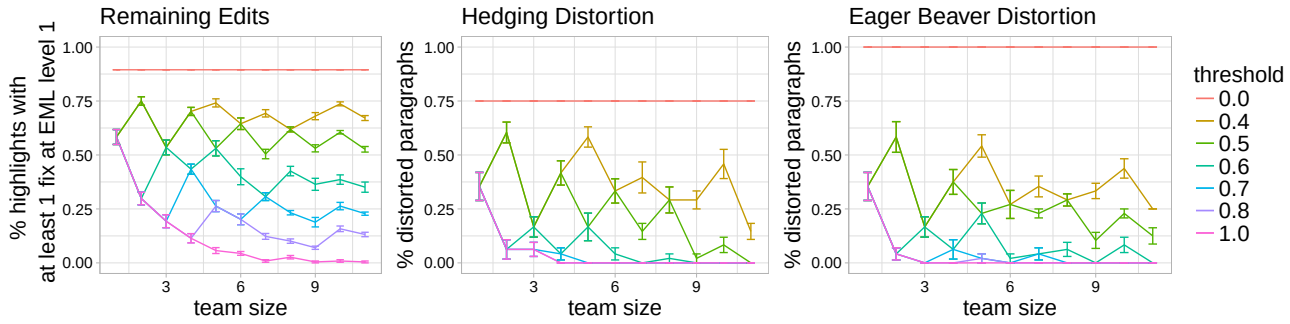

Figure 3. Results for the filter step. Increasing the agreement threshold reduces the number of paragraphs with a distorted suggestion (middle and right), but as a consequence, it also reduces the number of edits that make it through the system (left chart). For "eager beaver distortion" we allow one distortion since it only has a minor effect on the false positive rate, but setting the threshold high enough would eliminate even this error.

information while removing EML. These highlights would be better suited for a fix that restructures the sentence. Most highlighted portions landed in the third category, requiring clever rephrasing to remove EML while maintaining IEC.

Fix. In the second step we ask workers to suggest possible edits to remove EML from the highlighted portions in the previous step. We ask workers to provide an edit for all highlighted portions, scaling their pay based on the number of edits we ask them to make. Specifically, we provide the instructions: "Remove dramatic words and phrases from each of the highlighted portions while maintaining the same information and grammatical correctness." To prevent workers from attempting to fix multiple highlighted portions in the same edit, we restrict the range of text that workers are allowed to edit to include text starting from the end of the preceding highlight (or beginning of the text if the first highlight) to the start of the next highlight (or end of the text if the last highlight). While we give workers the option to skip highlights that are too challenging to rephrase, they generally provide a suggestion for all highlighted portions anyway.

In our component study, we find that performance plateaus at about 5 workers. After fixing this parameter we observe that 75% of highlighted portions will have at least one suggested edit that correctly removes all EML from the portion.

Filter. Our third step is to filter out suggested edits that distort the information of the original text. As we have noted above, distortion error can cause false positives later in the pipeline as workers conflate the change in information to be a change in the amount of EML. For IEC text, increased distortion can lead to increased conflation error.

Therefore, we build this component to include an *agreement threshold* that can be used to control the level of confidence that suggested edits passing through the system are not distorted. We ask workers to select suggested edits that "maintain the same information as the [original text]" and to "not select [suggested edits] that attempt to debias or soften the opinions in the [text]", then we aggregate suggestions based on the percentage agreement between workers. Setting the agreement threshold higher allows fewer edits through, but at higher confidence they are not distorted. Setting it lower may allow a higher percentage through, but without as much confidence in their quality.

In our study of this component, we noticed two ways that workers distort information in the original text: 1) as a result of workers' attempts to soften the opinions in the text by hedging the claims (we call this *hedging distortion*) and 2) as a result of workers too eagerly removing inconsequential information in addition to EML (we call this *eager beaver distortion*). We notice that hedging distortion can have a large effect on downstream classification, but eager beaver distortion less so (Figure 5). Despite this, we observe that nearly all hedging distortion and most eager beaver distortion can be eliminated with a 50% agreement threshold at 5 workers while ensuring that 50% of highlighted portions have a high quality suggestion (Figure 3). Setting the threshold to 80% eliminates all distortion while ensuring that 25% of highlighted portions have a high quality edit.

Select. The final step involves selecting the suggested edits that best remove EML from each highlighted portion. Of the suggested edits that make it through the filter in the previous step, we ask workers to "determine how well [each suggested edit] removes dramatic words and phrases from the highlighted portion of the [text] while still making sense" on a 3-point scale labeled "best", "decent", and "worst". For each highlighted portion, we select the suggestion that has the lowest mean score and incorporate the change into the final paragraph. Our component study found that 5 workers is sufficient for consistently selecting the best edits out of a group.

Iteration. We find that anchors can be iteratively improved by sending the output of the "Select" step back into "Find" step, reducing the EML level with each iteration. In our component study we found that iterations reduced the EML level of the anchor by an average of 1 point (of 4) with each iteration, eventually converging to an EML level of 1 after 3 iterations.

Comparison

After an anchor has been created in the anchor transformation step, we ask workers to assess the difference in EML between the anchor and the original text. Part of the contribution of our approach is that this step enables for the high-level task to be decomposed into highlight-level units. The comparison step aggregates these individual decisions into a final classification

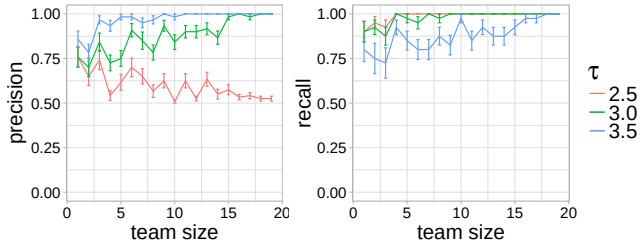

Figure 4. Precision and recall for the comparison step. "τ" indicates different decision boundaries on a 5-point Likert scale.

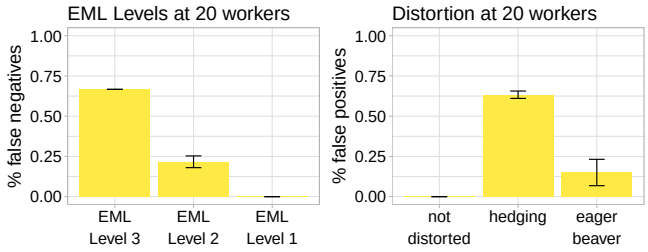

Figure 5. Errors in the anchor transformation step can compound into classification errors in the comparison step. Higher EML levels in the anchor cause an increased false negative rate and adding distortion leads to an increased false positive rate.

decision. We show workers the two versions of the text side-by-side and ask them to rank on a 5-point Likert scale "How much more dramatic is the [original text] compared to the [anchor]?" We aggregate ratings by taking the mean value and then by using a decision boundary to determine whether the text is to be classified as a positive or negative example.

Our evaluation of this component found that classification accuracy is dependent on the quality of the anchor used for comparison (Figure 4). For an anchor that correctly removes all EML from the text, 90% accuracy can be achieved with 5 workers and using more leads to even higher accuracy. However, we find that imperfect anchors cause classification error (Figure 5). For example, anchors that fail to completely remove EML (EML level greater than 1) are more likely to cause false negative classification because the difference in EML is perceived to be lower. Likewise, distortion in the anchor can create false positive classification because changes in information can be perceived as a difference in EML.

Putting It Together

Through a series of component studies, we have demonstrated the feasibility of anchor comparison. We showed that the anchor transformation process (left box in Figure 2) can create a version of the text with low EML and no distortion when a high agreement threshold and iteration are used. We also showed that the comparison process (right box in Figure 2) can achieve perfect precision and recall given the original text and a low EML version without distortion.

We will now walk through a version of our system that can achieve reasonable precision and recall. Given a set of new snippets for classification, we would first send them through the find step. On each iteration through anchor transformation (left box in Figure 2), we can first use 5 workers with a 20% overlap threshold, to highlight 75% of the EML words. Secondly, the snippets would go through the fix step. Using 5 workers, we can get high-quality suggestions for 75% of the highlighted portions. Thirdly, we send the snippets to the filter step. Our data shows that we can use 10 workers with a 50% agreement threshold to ensure 100% of paragraphs have no hedging error and 87% of paragraphs have at most one eager beaver error. However, we note that the line trends down (right graph in Figure 3), indicating that including more workers would feasibly ensure 100% of paragraphs have at most one eager beaver error. While distortion trends down, we note that the % of highlighted portions with a high quality suggestion remains constant (left graph in Figure 3). With three iterations through these steps (Find, Fix, Filter, Select), we would reach a final version. Finally, we would send our snippets through the compare step. Given that there is at most one eager beaver error in each snippet, we would be able to ensure perfect recall and 87% precision. System builders can also save costs by tuning classification errors for their needs.

While cost depends on the number of edits needed to remove EML, our data had an average of 5 highlighted portions per iteration, leading to a projected cost of about $47 per snippet. While this may be feasible for some applications, future work will need to address the engineering of reducing costs.

Tradeoffs

Our system affords builders two mechanisms to tune precision and recall. The first mechanism is *iteration* which can be used to control recall. Secondly, the *agreement threshold* can be used to control precision.

Iteration improves recall, but is more costly. The higher the EML level in the anchor text, the less the perceived difference between the original text and the anchor, and the more likely text with EML will be classified as having no EML. Iterations can be used to reduce the average EML level of anchors, and thereby reduce the false negative rate. However, increasing the number of iterations increases cost as more labor is required to find, fix, filter, and select edits.

Increasing the agreement threshold improves precision, but reduces efficiency. Allowing distortion in the anchor introduces the possibility that the distortion will be perceived as a difference in EML during the comparison step, leading to text with no EML to be classified as having EML. Increasing the agreement threshold can reduce the % of paragraphs with distortion error and reduce the false positive rate. However, increasing the agreement threshold has the side effect of reducing the number of suggestions that make it through the system, leading to a lower efficiency at which EML is extracted and causing an increased number of iterations needed to keep recall constant.

DISCUSSION

In this work, we examined the challenges associated with classifying text for social references. A specific challenge that makes that classification difficult is that references are often entangled with content, which leads to the potential for one to be misconstrued for the other. We have proposed a

new crowdsourcing approach that controls for conflation by transforming the classification problem into a comparison.

We then demonstrated the feasibility of our approach by exploring an important sub-problem that involves disentangling social references evoked through emotionally manipulative language (EML) from intrinsically emotional content (IEC). To test the limitations of existing approaches and the feasibility of our approach, we developed an appropriate test dataset and used it to evaluate five baseline approaches as well as ours.

We showed that existing approaches for classification struggle to distinguish between EML and IEC. Automatic approaches require substantial training data to understand social references which is both expensive to obtain and still leaves them vulnerable to novel language patterns that have not yet been added to training data. Crowdsourcing approaches perform better, but are prone to conflation error because non-expert human annotators struggle to disentangle sources of emotion in text —whether it is EML or IEC.

Our approach, anchor comparison, overcomes these challenges by leveraging workers' ability to find specific instances of EML in text and draft edits that remove them, resulting in an "anchor text" that can be used as a point with which to classify the original content by comparing the two. Through a series of component studies, we demonstrated that this approach is feasible and that it affords two mechanisms systems builders can use to improve precision and recall. At the penalty of increasing the cost of classification, iteration can be used to improve recall and an agreement threshold to improve precision.

Our contributions are, then, a class of problems that involve disentangling social reference from content, our anchor comparison approach that leverages transformation for disentangling these references, a system we created that uses anchor comparison, and an evaluation of our system on a dataset.

Limitations. However, our work has several limitations that remain for future work. First, the social references we address here assume that multivalent messaging is received by a homogeneous audience. Violations of this assumption (e.g., dog-whistles) have references for only sub-audiences and may need a specialized crowd to detect.

Second, we demonstrated through a proof-of-concept that anchor comparison can be used to detect EML without conflation error at a cost of about $50 given a short text snippet and sufficient run-time. In principle, this approach could be used to detect emotionally manipulative language online content. However, wide-scale deployment will require reducing cost. This remains for future work but we believe that this will be possible by combining machine learning approaches with the crowd. For example, one could use our crowdsourcing approach to generate a labeled dataset, which could in turn be used to train a classifier. Additionally, anchor transformation is particularly suitable for a hybrid intelligence approach, where machine learning is used to find and suggest potential edits to remove EML from the text, and crowd workers are used to make final judgments.

Third, while we took steps to ensure validity of our test dataset, our current implementation requires text snippets to be short and have all necessary context included in each snippet. Data encountered in the wild will likely include visual aspects, vary in length, and interconnected contexts. Future work will consider summarizing long news articles etc. that include all necessary context. Additionally, we may need to augment anchor transformation to include the crowd's capabilities with cropping or video editing tools.

Finally, while our study has considered one type of error (i.e., conflation error), we cannot draw firm conclusions about other kinds of error patterns that will occur in deployment settings. While we have found no evidence of alternative error patterns, we believe that political, racial, and other biases within the crowd may skew detection results. Prior work have proposed several strategies for mitigating these biases including [35, 69]. Future work should explore applying these strategies in tandem with anchor comparison to control multiple errors.

Despite these limitations, we remain cautiously optimistic. We have tackled one of the key challenges that makes media manipulation challenging to detect (conflation error) and introduced the first approach that can overcome this challenge. While there are likely many more challenges to overcome before our system can be robustly deployed at scale, we believe our fundamental approach is both feasible and important.

CONCLUSION

In this paper, we have contributed a new crowdsourcing approach by transforming a classification problem into a comparison. This allows the crowd to detect text that uses manipulative emotional language to sway users towards positions or actions. Our approach, *anchor comparison*, overcomes the challenges that cause automatic and standard crowdsourcing approaches for this problem to perform poorly: the difficulty of gathering comprehensive training data for social references and a tendency for the crowd to conflate emotionally manipulative language (EML) and intrinsically emotional content (IEC). We showed that our anchor comparison approach mitigates conflation errors by transforming the problem into a comparison task where IEC cannot overpower EML. We evaluated our approach by developing a corpus of short text pieces and also showed that our approach affords system builders the ability to tune precision and recall. We argue that our approach could be useful for identifying potential-manipulative content and warning users, helping the public see and understand media manipulation. More generally, our approach is a first step toward solving problems that involve disentangling social references from content.

ACKNOWLEDGMENTS

This research was supported in part by the U.S. NASA (NNX16AC66A) and DARPA (D19AP00079) agencies. This article solely reflects the opinions and conclusions of its authors and not NASA, DARPA, or any other government entity. We also thank the members of the SocialWorlds Research Group and Cromalab.

REFERENCES

[1] Rama Akkiraju. 2015. IBM Watson Tone Analyzer–new service now available. *IBM Cloud Blog, Jul* 16 (2015).

[2] Aristotle. *Rhetorica*. 1:1.

[3] Fatemeh Torabi Asr and Maite Taboada. 2018. The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 10–15.

[4] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313–322.

[5] Monika Bickert. 2019. Combatting Vaccine Misinformation. *Facebook* (2019).

[6] Herbert Bless, Gerald L Clore, Norbert Schwarz, Verena Golisano, Christina Rabe, and Marcus Wölk. 1996. Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of personality and social psychology* 71, 4 (1996), 665.

[7] Galen V Bodenhausen, Lori A Sheppard, and Geoffrey P Kramer. 1994. Negative affect and social judgment: The differential impact of anger and sadness. *European Journal of social psychology* 24, 1 (1994), 45–62.

[8] Nadia M Brashier, Emmaline Drew Eliseev, and Elizabeth J Marsh. 2020. An initial accuracy focus prevents illusory truth. *Cognition* 194 (2020), 104054.

[9] Alan Brinton. 1988. Pathos and the" Appeal to Emotion": An Aristotelian Analysis. *History of Philosophy Quarterly* 5, 3 (1988), 207–219.

[10] Kristie Canegallo. 2019. Fighting disinformation across our products. *Google* (2019).

[11] Robyn Caplan, Lauren Hanson, and Joan Donovan. 2018. Dead Reckoning Navigating Content Moderation After “Fake News”. *Data & Society* (2018).

[12] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 39–48.

[13] David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*. ACM, 128–135.

[14] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Daniel S Weld. 2018. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. *arXiv preprint arXiv:1810.10733* (2018).

[15] John J.Y. Chung, Jean Y. Song, Sindhu Kutty, Sungsoo Ray Hong, Juho Kim, and Walter S. Lasecki. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. In *Proceedings of the ACM conference on Computer-Supported Collaborative Work (CSCW ’19)*. ACM, New York, NY, USA. DOI: http://dx.doi.org/10.1145/3359164

[16] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, and others. 2019. Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior* (2019), 1–23.

[17] John Cook, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one* 12, 5 (2017), e0175799.

[18] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 519–535.

[19] Krista De Castella, Craig McGarty, and Luke Musgrove. 2009. Fear appeals in political rhetoric about terrorism: An analysis of speeches by Australian Prime Minister Howard. *Political Psychology* 30, 1 (2009), 1–26.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[21] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[22] Damer T Edward. 2008. Attacking Faulty Reasoning: A Practical Guide to Fallacy-free Arguments. *Cengage Learning* 209 (2008).

[23] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining.. In *LREC*, Vol. 6. Citeseer, 417–422.

[24] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4647–4657.

[25] Melissa L Finucane, Ali Alhakami, Paul Slovic, and Stephen M Johnson. 2000. The affect heuristic in judgments of risks and benefits. *Journal of behavioral decision making* 13, 1 (2000), 1–17.

[26] Roger Fowler. 2013. *Language in the News: Discourse and Ideology in the Press*. Routledge.

[27] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[28] Michael Golebiewski and danah boyd. 2018. Data Voids: Where Missing Data Can Easily Be Exploited. *New York: Data & Society Research Institute* (2018).

[29] Mitchell Gordon, Jeffrey P Bigham, and Walter S Lasecki. 2015. LegionTools: a toolkit+ UI for recruiting and routing crowds to synchronous real-time tasks. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 81–82.

[30] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 534.

[31] Vladas Griskevicius, Noah J Goldstein, Chad R Mortensen, Jill M Sundie, Robert B Cialdini, and Douglas T Kenrick. 2009. Fear and loving in Las Vegas: Evolution, emotion, and persuasion. *Journal of Marketing Research* 46, 3 (2009), 384–395.

[32] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council* 9 (2018).

[33] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[34] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2018. LimitBias! Measuring Worker Biases in the Crowdsourced Collection of Subjective Judgments. (2018).

[35] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 407.

[36] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[37] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 64–67.

[38] Youxuan Jiang, Catherine Finegan-Dollak, Jonathan K Kummerfeld, and Walter Lasecki. 2018. Effective Crowdsourcing for a New Type of Summarization Task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Vol. 2. 628–633.

[39] Eric J Johnson and Amos Tversky. 1983. Affect, generalization, and the perception of risk. *Journal of personality and social psychology* 45, 1 (1983), 20.

[40] Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review* 93, 5 (2003), 1449–1475.

[41] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. ACM, 1003–1012.

[42] Richard S Lazarus. 1991. Cognition and motivation in emotion. *American psychologist* 46, 4 (1991), 352.

[43] Jennifer S Lerner, Roxana M Gonzalez, Deborah A Small, and Baruch Fischhoff. 2003. Effects of fear and anger on perceived risks of terrorism: A national field experiment. *Psychological science* 14, 2 (2003), 144–150.

[44] Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. 2015. Emotion and decision making. *Annual review of psychology* 66 (2015).

[45] Rebecca Lewis. 2018. Alternative Influence: Broadcasting the Reactionary Right on YouTube. *New York: Data & Society Research Institute* (2018).

[46] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, 142–150.

[47] VK Chaithanya Manam and Alexander J Quinn. 2018. Wingit: Efficient refinement of unclear task instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[48] Beth Innocenti Manolescu. 2006. A normative pragmatic perspective on appealing to emotions in argumentation. *Argumentation* 20, 3 (2006), 327–343.

[49] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute* (2017).

[50] Alice E Marwick. 2018. Why Do People Share Fake News? A Sociotechnical Model of Media Effects. *Georgetown Law Technology Review* (2018).

[51] Hana Matatov, Adina Bechhofer, Lora Aroyo, Ofra Amir, and Mor Naaman. 2018. DejaVu: A System for Journalists to Collaboratively Address Visual Misinformation. In *Computation + Journalism Symposium*. Miami, FL. `https://scholar.harvard.edu/files/oamir/files/dejavu-system-journalists_3.pdf`

[52] Arunesh Mathur, Gunes Acar, Michael Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *arXiv preprint arXiv:1907.07032* (2019).

[53] Katerina E Matsa and Shearer Elisa. 2018. News Use Across Social Media Platforms 2018. *Pew Research Center* (Sept 2018). http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

[54] David Merritt, Jasmine Jones, Mark S Ackerman, and Walter S Lasecki. 2017. Kurator: Using the crowd to help families with personal curation tasks. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1835–1849.

[55] Raphaël Micheli. 2010. Emotions as objects of argumentative constructions. *Argumentation* 24, 1 (2010), 1–17.

[56] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1345–1354.

[57] Tanushree Mitra, Graham Wright, and Eric Gilbert. 2017a. Credibility and the dynamics of collective attention. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 80.

[58] Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017b. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 126–145.

[59] Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, 26–34.

[60] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[61] Adam Mosseri. 2016. Addressing Hoaxes and Fake News. *Facebook* (2016).

[62] Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355* (2019).

[63] Gordon Pennycook and David G Rand. 2017. The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. (2017).

[64] Gordon Pennycook and David G Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* (2018).

[65] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104* (2017).

[66] Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1403–1412.

[67] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. 2011. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing*. IEEE, 3097–3100.

[68] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 502–518.

[69] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[70] Elizabeth A Segal. 2018. *Social Empathy: The Art of Understanding Others*. Columbia University Press.

[71] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[72] Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology* 48, 4 (1985), 813.

[73] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.

[74] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[75] Melodie Yun-Ju Song and Anatoliy Gruzd. 2017. Examining Sentiments and Popularity of Pro-and Anti-Vaccination Videos on YouTube. In *Proceedings of the 8th International Conference on Social Media & Society*. ACM, 17.

[76] Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *Eleventh International AAAI Conference on Web and Social Media*.

[77] Luc Steels and Manfred Hild. 2012. *Language grounding in robots*. Springer Science & Business Media.

[78] Carlo Strapparava, Alessandro Valitutti, and others. 2004. Wordnet affect: an affective extension of wordnet.. In *Lrec*, Vol. 4. Citeseer, 40.

[79] Larissa Z Tiedens and Susan Linton. 2001. Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing. *Journal of personality and social psychology* 81, 6 (2001), 973.

[80] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[81] Timothy D Wilson and Nancy Brekke. 1994. Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin* 116, 1 (1994), 117.

[82] Meng-Han Wu and Alexander James Quinn. 2017. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

[83] Anbang Xu and Brian Bailey. 2012. A reference-based scoring model for increasing the findability of promising ideas in innovation pipelines. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1183–1186.

[84] Jason R Young. 2003. The role of fear in agenda setting by television news. *American Behavioral Scientist* 46, 12 (2003), 1673–1695.

[85] Muhammad Bilal Zafar, Parantapa Bhattacharya, Niloy Ganguly, Saptarshi Ghosh, and Krishna P Gummadi. 2016. On the wisdom of experts vs. crowds: discovering trustworthy topical news in microblogs. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 438–451.

[86] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326* (2018).

[87] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, and others. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 603–612.

[88] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*. 649–657.