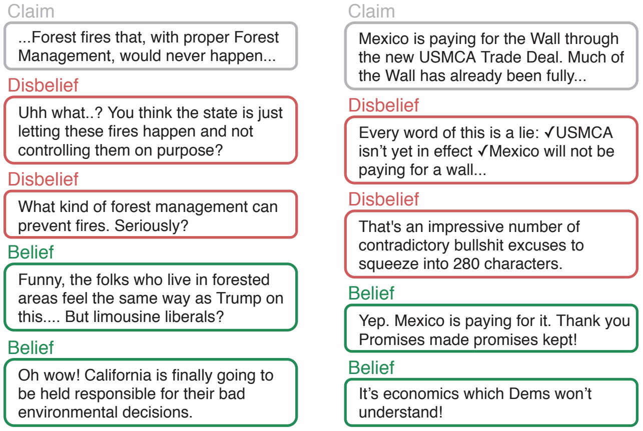
Modeling and Measuring Expressed (Dis)belief in (Mis)information

Shan Jiang,[♠] **Miriam Metzger,**[♣] **Andrew Flanagin,**[♣] **Christo Wilson**[♠]
[♠]Northeastern University, [♣]UC Santa Barbara
{sjiang, cbw}@ccs.neu.edu, {metzger, flanagin}@ucsb.edu

Abstract

The proliferation of online misinformation has been raising increasing societal concerns about its potential consequences, e.g., polarizing the public and eroding trust in institutions. These consequences are framed under the public's susceptibility to such misinformation — a narrative that needs further investigation and quantification. To this end, our paper proposes an observational approach to model and measure expressed (dis)beliefs in (mis)information by leveraging social media comments as a proxy. We collect a sample of tweets in response to (mis)information and annotate them with (dis)belief labels, explore the dataset using lexicon-based methods, and finally build classifiers based on the state-of-the-art neural transfer-learning models (BERT, XLNet, and RoBERTa). Under a domain-specific thresholding strategy for unbiasedness, the best-performing classifier archives macro-F_1 scores around 0.86 for disbelief and 0.80 for belief. Applying the classifier, we conduct a large-scale measurement study and show that, for true/mixed/false claims on social media, 12%/14%/15% of comments express disbelief and 26%/21%/20% of comments express belief. In addition, our results suggest an extremely slight time effect of falsehood awareness, a positive effect of fact-checks to false claims, and differences in (dis)belief across social media platforms.



Claim

...Forest fires that, with proper Forest Management, would never happen...

Disbelief

Uhh what..? You think the state is just letting these fires happen and not controlling them on purpose?

Disbelief

What kind of forest management can prevent fires. Seriously?

Belief

Funny, the folks who live in forested areas feel the same way as Trump on this.... But limousine liberals?

Belief

Oh wow! California is finally going to be held responsible for their bad environmental decisions.

(a) The **CA wildfire** claim and tweets expressing (dis)belief.

Claim

Mexico is paying for the Wall through the new USMCA Trade Deal. Much of the Wall has already been fully...

Disbelief

Every word of this is a lie: ✓USMCA isn't yet in effect ✓Mexico will not be paying for a wall...

Disbelief

That's an impressive number of contradictory bullshit excuses to squeeze into 280 characters.

Belief

Yep. Mexico is paying for it. Thank you Promises made promises kept!

Belief

It's economics which Dems won't understand!

(b) The **Mexico wall** claim and tweets expressing (dis)belief.

Figure 1: Example comments expressing (dis)belief in response to two false claims made on Twitter.

1 Introduction

Misinformation, broadly defined as any false or inaccurate information, has been spreading epidemically on social media (Lazer et al. 2018). During the 2016 US presidential election cycle, researchers estimated that "fake news" accounted for 6% of all news consumption (Grinberg et al. 2019), and 44% of Americans age 18 or older visited at least one untrustworthy website (Guess, Nyhan, and Reifler 2018). To date, misinformation has been documented across the globe, e.g., in Africa (Wasserman and Madrid-Morales 2019), Asia (Kaur et al. 2018), and Europe (Fletcher et al. 2018).

The proliferation of misinformation has been raising increasing societal concerns about its potential consequences. In the political context, fabricated stories and partisan opinions may polarize the public (Levendusky 2013), alter voters' perceptions about candidates (Allcott and Gentzkow 2017; Epstein and Robertson 2015), and erode trust in institutions (Ciampaglia et al. 2018), therefore posing a threat to the democracy (Morgan 2018; Hochschild and Einstein 2015).

These consequences are framed under the public's susceptibility to misinformation, as the public is unable, or disinclined, to distinguish truth from fiction. This narrative, however, needs further investigation and quantification. Recent surveys from the Reuters Institute and Pew Research Center reported that the public is indeed aware of the misinformation problem, and (dis)believes certain information sources (e.g., news outlets, politicians) more than others (Anderson and Rainie 2017; Nielsen and Graves 2017). However, these studies are small-scale in nature, and thus unable to quantitatively measure *to what extent does the public (dis)believe in (mis)information*.

Complementary to these surveys, our work proposes an observational approach as an alternate lens through which to interrogate the public's (dis)belief in (mis)information. Our approach leverages social media comments as a proxy for assessing individuals' responses to (mis)information. Consider the examples shown in Figure 1: the language used in comments in response to claims can express signals of the users' (dis)belief, therefore, if modeled properly, these social media comments can be used to measure the prevalence of *expressed* (dis)belief at scale.[1]

The first part of this paper explores methods to **model** (dis)belief expressed in comments. We start by collecting a small sample of tweets that comment on fact-checked claims, and then manually annotate each tweet with disbelief and belief labels. Using this dataset, we experiment with Natural Language Processing (NLP) models. We first conduct an exploratory analysis using lexicon-based methods, which reveals differences in word usage (e.g., falsehood awareness signals, positive and negative emotions) in tweets expressing (dis)belief verses others. Next, we experiment with classification models, including linear models with lexicon-derived features, as well as state-of-the-art neural transfer-learning models (e.g., BERT (Devlin et al. 2019), XLNet (Yang et al. 2019), and RoBERTa (Liu et al. 2019)). Then, we develop a domain-specific thresholding strategy for classifiers to make unbiased predictions compared to human experts. Under chosen thresholds, the best-performing classifier achieves macro-F_1 scores around 0.86 for predicting disbelief and 0.80 for belief. We have released our data, code, and trained models to accelerate the development of future studies.[2]

The second part of the paper aims to **measure** expressed (dis)belief at scale by applying the trained classifier. We run the classifier on an existing large, unlabeled dataset of social media comments in response to (mis)information (collected during our prior work (Jiang and Wilson 2018)) and analyze the estimated prevalence of expressed (dis)belief. Our results show that:

- **RQ1**, *overall prevalence*: For true/mixed/false claims on social media, 12%/14%/15% of comments express disbelief and 26%/21%/20% of comments express belief, suggesting (optimistically) increased disbelief and decreased belief as information veracity decrease, yet (pessimistically) considerable suspicions on truthful information;
- **RQ2**, *time effect*: There is an extremely slight time effect of falsehood awareness, where disbelief increases 0.001% and belief decreases 0.002% per day after a false claim is published;
- **RQ3**, *fact-check effect*: Controlling for the time effect, disbelief increases 5% and belief decreases 3.4% after claims are fact-checked, suggesting a positive effect of fact-checks on altering the prevalence of (dis)belief;
- **RQ4**, *platform differences*: There is a difference in (dis)belief expressed across three mainstream social media platforms (Facebook, Twitter, and YouTube).

In the rest of the paper, § 2 positions our work within related areas, § 3 describes the data collection and annotation process, § 4 experiments with NLP methods to model expressed (dis)belief, § 5 applies the model to measure expressed (dis)belief and answers (**RQ1–4**), and finally § 6 discusses the limitations of this observational approach and potential directions for future work.

2 Related Work

Our work to model and measure expressed (dis)belief in (mis)information connects a rich line of literature. This section positions our work within related misinformation studies, and discerns our work from related NLP tasks and datasets.

2.1 (Dis)belief and (Mis)information

There is an emerging line of work focusing on the misinformation topic, ranging from its political influence (Allcott and Gentzkow 2017; Grinberg et al. 2019; Guess, Nagler, and Tucker 2019; Robertson et al. 2018; 2019) to algorithmic detection (Shu et al. 2017; Zhou et al. 2019) and intervention (Jiang et al. 2020; Jiang, Robertson, and Wilson 2020; 2019; Farajtabar et al. 2017; Jang and Kim 2018). Some of these studies adopted the recent (yet controversial) term "fake news", while we choose to use the term "misinformation" as it covers a broader spectrum of information veracity (e.g., partial truths, as opposed to blatant lies), and is not as politicized as "fake news".

A key question in this literature is (**RQ1**) *does the public believe misinformation, and if so, to what extent?* If misinformation is not believed, that would discount much of its alleged influence on the public and the political process.

Insights into (**RQ1**) are provided by existing psychological and sociological theories that hypothesize about the public's susceptibility to misinformation. Naïve realism (Ward et al. 1997) and confirmation bias theory (Nickerson 1998) from psychology suggested that people tend to believe in information that resonates with their pre-existing (yet potentially false) beliefs. Social identity (Stets and Burke 2000) and normative influence theory (Kincaid 2004) from sociology suggested that people tend to follow the norms of their established ideological groups when responding to information, and spread their beliefs in "socially safe" information, often regardless of its veracity.

On the empirical side, a report from the Pew Research Center provided evidence for these theories by conducting a survey about trust in news outlets across the ideological spectrum. It found a significant correlation between **(a)** the self-reported trust and **(b)** the ideological proximity between the audience and the news outlet, e.g., the liberal audience tended to trust the New York Times while conservative audiences did not, and vice-versa for Fox News (Mitchell et al. 2014). More recent reports from the Reuters Institute (Nielsen and Graves 2017) and Pew Research Center (Anderson and Rainie 2017) surveyed in more depth about the socio-psychological mechanisms behind (dis)belief and (mis)information, and reported that the public is indeed aware of the misinformation problem. Despite the valuable evidence they offered, these qualitative and experimental studies are small-scale, and they required direct interactions with the participants, therefore potentially suffering from the Hawthorne Effect where participants modified their behaviors under their awareness of being surveyed (McCarney et al. 2007).

[1] We discuss the emphasis on *expressed* in § 6.1.

[2] Available at: https://misinfo.shanjiang.me

Quantitative research on this topic is relatively limited. (Jiang and Wilson 2018) analyzed social media comments in response to misinformation using an unsupervised approach, and showed that certain linguistic signals suggesting (dis)belief (e.g., "fake", "dumbest") were distributed differently in response to claims with differing veracity. In § 4.1, we verify that these signals do indeed correlate with the likelihood to express (dis)belief, but they are insufficient predictors to judge if a comment expresses (dis)belief.[3]

In addition to (**RQ1**), we also investigate two follow-up research question. The first is (**RQ2**) *if there is a time effect* for expressed (dis)belief in misinformation, where the public gradually realizes the truth after a claim is made, and therefore loses trust in false claims over time. This question is raised in light of recent work that leverages the "wisdom of the crowd" for misinformation detection (Tschiatschek et al. 2018; Kim et al. 2018).

The second question is (**RQ3**) *if fact-checks have an effect* on expressed (dis)belief, after a false claim is judged by a certified fact-checker (e.g., Snopes, PolitiFact) (Poynter 2020). This question continues an ongoing debate on the importance, or lack thereof, of fact-checking in the misinformation ecosystem (Tambuscio et al. 2015; Garrett, Nisbet, and Lynch 2013).

2.2 Related NLP Tasks and Datasets

In the realm of computational social science, automatically *scoring* a dataset is a common prerequisite for hypothesis testing. Existing studies that used language as a signal mostly adopted a simple, straightforward scoring method that leveraged unigram-based bag-of-words (BoW) models (Gentzkow, Kelly, and Taddy 2019; Hu et al. 2019). This method, however, could have limited applicability for our task, as identifying expressed (dis)belief could be more subtle than their tasks (e.g., identifying ideology), and therefore requires models to comprehend entire statements as a whole instead of averaging signals of unigrams.

In the realm of NLP, however, such *scoring* is the native output of probabilistic classifiers, and the above method is equivalent to linear models with BoW features on a sequence classification problem (Xing, Pei, and Keogh 2010). More recent and better solutions for this task use neural architectures (Lai et al. 2015; Zhou, Wan, and Xiao 2016) and pre-trained transfer-learning models (Devlin et al. 2019; Yang et al. 2019; Liu et al. 2019).

Specific applications of the sequence classification problem are defined within domain-specific datasets. Although there is, to our knowledge, no existing dataset on detecting (dis)belief, there are proposed NLP tasks that are related to our task. Stance detection, for example, aims to determine the for-or-against stance in comments for a two-sided argument (e.g., marijuana, gay marriage) (Hasan and Ng 2013; Joseph et al. 2017), and, in the political context, it often overlaps with ideology identification (Preoţiuc-Pietro et al. 2017). Intuitively, this task is similar to our problem, however, we do observe conflicting cases in our data where comments sharing the same ideological stance provide informative counter-evidence on the claims and therefore express disbelief.

Similarly, classifications of other creative languages such as sarcasm (González-Ibánez, Muresan, and Wacholder 2011), satire (Burfoot and Baldwin 2009), irony (Farías, Patti, and Rosso 2016), and humor (Yang et al. 2015) share certain commonalities with our task, but none of them fulfills our need to identify (dis)belief.

3 Data

To model (dis)belief in (mis)information, we first collect a sample of tweets written in response to claims that are potentially misinformation, and then manually annotate them with belief and disbelief labels.

3.1 Collecting Comments on Misinformation

Finding (mis)information claims. PolitiFact is an IFCN-certified fact-checking agency that evaluates the veracity of wide-spread claims online (Poynter 2020). We read through all of PolitiFact's fact-check articles written between January 1 to June 1, 2019 and manually found the ones whose claims originated from Twitter. We recorded the IDs of the tweets containing these claims.

Collecting comments in response to claims. Using the above fact-checked tweets as seeds, we queried an archived 1% sample of the tweet stream (Liu, Kliman-Silver, and Mislove 2014) and found all *comments* to the seed tweets. In Twitter's terminology, these comments include "replies" and "retweets with comments" (i.e., quoted tweets) but excludes other retweets (Twitter 2020). Note that we only keep comments whose text content is non-empty, as we aim to identify expressed (dis)belief using language features.

To filter out noise, we keep only the claims that we could link to >50 comments, which resulted in 18 claims with 6,809 comments. The short names of these claims are displayed as the x-tick labels in Figure 2. The full description of each claim and corresponding fact-check articles is available in our published dataset.

Representativeness. Although our archived 1% sample of the tweet stream has been shown to be representative of the Twitter ecosystem as a whole (Morstatter et al. 2013), this dataset is *not* a representative sample to understand the prevalence of (dis)belief at scale. This is due to **(a)** the narrow time period (i.e., half a year) of seed claims and comments, and **(b)** the omission of other mainstream social media platforms (e.g., Facebook, YouTube). While **(a)** is a common limitation on longitudinal validity in the literature (Street and Ward 2012), **(b)** is less commonly considered. (Zannettou et al. 2017) reported that misinformation sharing behaviors differ across platforms, which motivates our last research question (**RQ4**) *if expressed (dis)belief is distributed differently across social media platforms*.

Taken together, these two issues mean that high-level statistics from this sample cannot be used to measure (dis)belief and test related hypothesis. Hence, we leverage a much larger dataset in § 5). However, this sample is useful to understand

[3] As some intuitive examples, the comments shown in Figure 1 have no obvious unigram signals signifying (dis)belief.

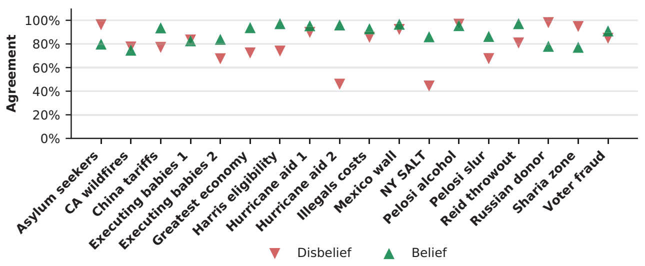


Figure 2: Inter-annotator agreement by claim. Out of 36 evaluated groups/labels, 66.7% are above 80% agreement and 88.9% are above 70% agreement.

the *language* that people used to express their (dis)belief in response to (mis)information.

3.2 Annotating (Dis)belief Labels

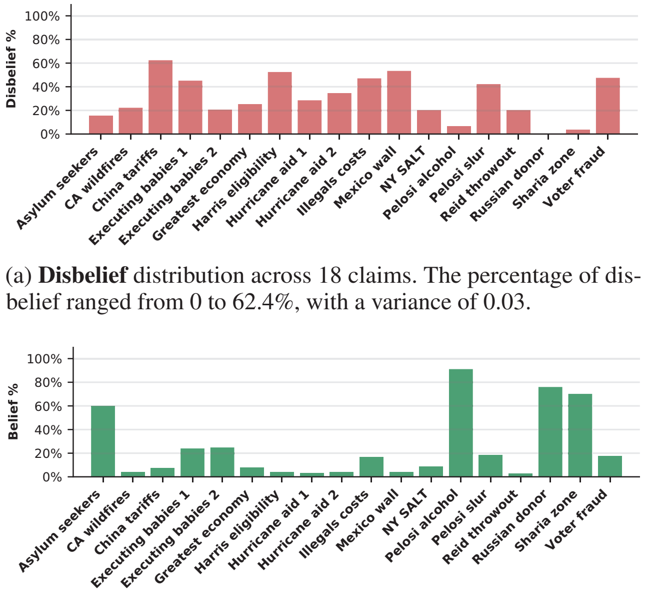
We annotate our unlabeled dataset of comments with belief and disbelief labels by recruiting a group of communication-majored undergrads and a faculty member from the communication department as the expert.

Task assignment. Annotating 6,809 tweets is a heavy task. To reduce the workload, we grouped these tweets by the initial claims and assigned each group of tweets to two independent human annotators. We trained the annotators, and then asked them to provide binary labels on each tweet in the given group: *disbelief* (i.e., if the person who wrote the comment *does not* believe the claim) and *belief* (i.e., if the person who wrote the comment *does* believe the claim). Note that these two labels are mutually exclusive but not necessarily complementary, i.e., we do not expect a tweet to show both belief and disbelief, but it can show neither.

Inter-annotator agreement. Our task assignment strategy allows us to evaluate inter-annotator agreement at the individual group level. We use the inter-annotator percent agreement[4] (i.e., the number of agreed labels over the total count) for each group and each label, and show the results in Figure 2. Out of 36 evaluated groups/labels, 66.7% (24/36) are above 80% agreement, 88.9% (32/36) are above 70% agreement, and only two are below 60% agreement, suggesting a high level of agreement among annotators, especially for a relatively subjective task.

Final labels. To obtain a final label for each tweet, a faculty member from the communication department read through all cases where two annotators disagreed and then provided a final judgement to break ties. This effectively makes our annotation process a majority vote among three members.

Note that there are two straightforward ways to formulate the (dis)belief labels: **(a)** a single-label quadruple-class formulation, where the four possible classes are: belief, disbelief, both, and neither; or **(b)** a double-label binary formulation, where one label is belief or not and the other is disbelief or not. Although these two formulations are equivalent here, **(b)** provides us with more flexibility for classification, as it is easy to threshold on each binary label and easy to analyze the performance tradeoff (as we discuss in § 4.2). Thus, we choose formulation **(b)** for the (dis)belief labels.



(a) **Disbelief** distribution across 18 claims. The percentage of disbelief ranged from 0 to 62.4%, with a variance of 0.03.

(b) **Belief** distribution across 18 claims. The percentage of belief ranged from 2.8% to 91.1%, with a variance of 0.08.

Figure 3: Data overview by claim. There is large variation in expressed (dis)belief across the 18 claims, and the distributions of disbelief and belief are negatively correlated.

Data overview. Overall, out of 6,809 tweets, 2,399 (35.2%) are labeled as expressing disbelief, 1,282 (18.8%) are labeled as expressing belief, 3,128 (45.9%) are labeled as neither and none (0%) are labeled as both. Disbelief is over-represented in this sample (cf. the overall prevalence measured in § 5.1) as the 18 claims in the sample contain heavy misinformation.

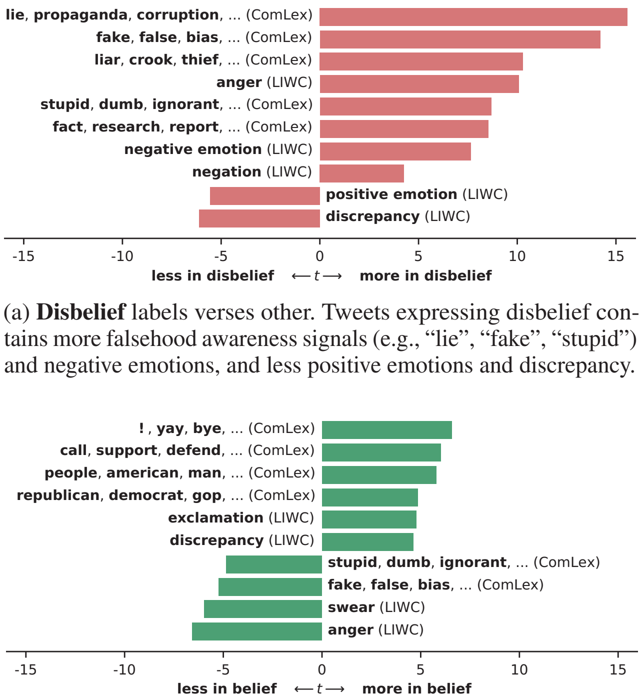
The distribution of (dis)belief for each claim is shown in Figure 3. There is large variation in expressed (dis)belief across the 18 claims, and the distributions of disbelief and belief are, as expected, negatively correlated (Pearson $r = -0.68^{***}$).[5]

4 Model

Leveraging our labeled dataset, we first conduct a lexicon-based exploratory analysis of language used across tweets expressing belief and disbelief, and then experiment with NLP models to build classifiers.

[4] Cohen's κ is not preferred here, as (dis)belief labels are, by our hypotheses, unevenly distributed and therefore κ's baseline agreement is irrelevant.

[5] $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.



(a) **Disbelief** labels verses other. Tweets expressing disbelief contains more falsehood awareness signals (e.g., "lie", "fake", "stupid") and negative emotions, and less positive emotions and discrepancy.

(b) **Belief** labels verses other. Tweets expressing belief contains more exclamation (e.g., "!", "yay") and discrepancy, and less falsehood awareness signals (e.g., "lie", "fake", "stupid") and negative emotions.

Figure 4: Language difference between tweets expressing (dis)belief and others. Significance of difference is obtained by t-tests with $p < 0.01$ after Bonferroni correction. Ten samples of significant categories of LIWC and ComLex with their respected t-values and category names are shown for both disbelief and belief labels.

4.1 Exploratory Analysis of Language

We start the modeling of (dis)belief by exploring the question *if tweets expressing (dis)belief use different language than the others, and if so, what are the differences?*

We adopt a lexicon-based method to explore this question, and choose two lexicons: **(a)** LIWC (Tausczik and Pennebaker 2010), the most widely-used lexicon for understanding psychometric properties of language, containing generic emotional and topical word categories, e.g., "anger", "reward", "work"; and **(b)** ComLex (Jiang and Wilson 2018), a more contextual lexicon built from social media comments to misinformation, containing additional domain-specific categories, e.g., "fake", "fact", "hate speech".

Each word category in the lexicon contains a set of curated words that embody signals of the category (e.g., "sad" for "negative emotion"). Briefly, our method works as follows: we apply a lexicon on a tweet, which results in a frequency f_c for each category c in the lexicon, counting the overlap between words in the tweet and words in the corresponding category c. Then, at the dataset level, we compare the distributions of such frequency between tweets expressing (dis)belief and the others, by performing independent t-test for $\mathbb{E}(f_c)$. Significance is obtained by setting $p < 0.01$ after Bonferroni correction on the number of categories (392 total categories: 92 for LIWC and 300 for ComLex). Ten representative samples of significant categories with their t-values and category names[6] are shown in Figure 4.

Figure 4a shows that tweets expressing disbelief contain more falsehood awareness signals, including referrals to falsehood "lie, propaganda, ..." ($t = 15.6^{***}$) and "fake, false, ..." ($t = 14.2^{***}$), referrals to the truth "fact, research, ..." ($t = 8.5^{***}$), and negative character portraits such as "liar, crook, ..." ($t = 10.3^{***}$) and "stupid, dumb, ..." ($t = 8.7^{***}$). These results are intuitive and provide face-validity to the existing linguistic study of misinformation responses, where similar signals were used to insinuate users' disbelief (Jiang and Wilson 2018). In addition, tweets expressing disbelief also contain more negative emotions ($t = 7.6^{***}$) and negation (e.g., "no, not", $t = 4.3^{***}$), less positive emotions ($t = -5.6^{***}$) and discrepancy (e.g., "should, would", $t = -6.1^{***}$).

Figure 4b shows that tweets expressing belief contain less falsehood awareness signals, including referrals to falsehood "fake, false, ..." ($t = -5.2^{***}$) and negative character portrait "stupid, dumb, ..." ($t = -4.8^{***}$). This is intuitively the opposite of disbelief. In addition, tweets expressing belief also contain more exclamation (for both LIWC exclamation marks, $t = 4.8^{***}$, and ComLex "!, yay, ..." category, $t = 6.6^{***}$) and discrepancy ($t = 4.6^{***}$), and less negative reactions such as swear (e.g., "damn, fuck", $t = -6.0^{***}$) and anger (e.g., "hate, kill", $t = -6.6^{***}$).

4.2 Experiments with Classification Models

Given these observed difference in language usage, our next question is *if such difference can be used to identify tweets that express (dis)belief?* To answer this question, we experiment with NLP models to build classifiers.

Chance. We first experiment with a chance classifier where we assign random probabilities for both disbelief and belief labels to demonstrate trivial performance baselines.

Lexicon-derived features with linear models. As a continuation of § 4.1, we run experiments using lexicon-derived features with linear models. For each tweet, we concatenate all mapped frequencies f_c across all categories c to a vector representation $\vec{f}$ (92 dimensions for LIWC and 300 for ComLex), and then feed these vector representations to a Logistic Regression (LR) layer for classification.

These models should perform better than trivial baselines, as they include the language signals we observed in § 4.1. However, their performance is still inherently limited, as such methods only capture the semantics of unigrams while ignoring the dependency between words (e.g., co-reference, phrases). Thus, these models are incapable of comprehending an entire tweet at the sequence level.

Neural transfer-learning models. To boost performance, we embed the entire sequence and leverage state-of-the-

[6] ComLex has some unnamed categories, in which case we use three words in that category as the category name.

art neural transfer-learning (Pan and Yang 2009) methods for the task. We experiment with three pre-trained models: BERT (Devlin et al. 2019), XLNet (Yang et al. 2019), and RoBERTa (Liu et al. 2019).

This method follows a *pre-training-fine-tuning* paradigm. During the *pre-training* phase, transformer (Vaswani et al. 2017) or transformer-XL (Dai et al. 2019) based models are trained on large, unlabeled corpus with certain objectives, e.g., BERT and RoBERTa are trained to predict missing words in sentences, XLNet is trained to predict last tokens in factorization orders of sentences. During this process, a randomly initialized model is adjusted by back-propagation of loss, and its weights are progressively updated to embed knowledge of human language.

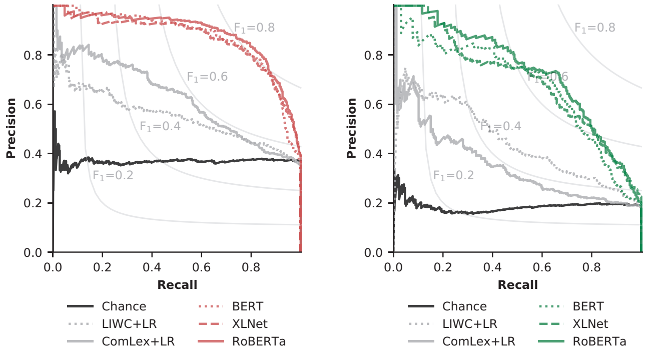
During the *fine-tuning* phase, models are initialized with pre-trained weights and then re-train on labeled data over specific tasks. This process tunes an already sophisticated model to perform specific downstream tasks, thus the model is expected to achieve high performance on a small labeled dataset.

To experiment with these neural models, we first preprocess tweets through the same pipeline designed in the pre-training phase, which includes tokenizing tweets at the sub-word level using specific tokenizer, and then padding or truncating the sequence to a specific length.[7] Next, these sequences are fed to an input layer which is connected to a pre-trained model. After all parameters flow through the model, we replace the last layer of the model with a double-label classification layer to predict (dis)belief. Finally, we compare the predictions and labels, calculate the cross entropy loss, and back-propagate errors. This training process is done iteratively for a certain number of epochs, as determined by cross validation on the training set.

As reported in the original papers, these models achieve state-of-the-art performance on a wide range of generic NLP tasks. Thus, we expect that they can increase performance for our task (versus the linear models) without designing domain-specific neural architectures.

Experimental setup. We randomly split the dataset into 80% (5,448) training set and 20% (1,361) testing set. Our linear models were trained until convergence, which completed within one minute. We set up the neural models (BERT, XLNet, and RoBERTa) using the same neural architecture, hyperparameters, vocabularies, and tokenizers as the base models described in the original papers,[8] and we trained them for three epochs, which completed within two hours on a single Titan X Pascal GPU.

Evaluation metrics. All of the models we experiment with are probabilistic classifiers that assign a probability $\mathbb{P}$ to the positive label (i.e., disbelief or belief) and the remaining $1 - \mathbb{P}$ to the negative label (i.e., not disbelief or not belief). We then obtain the predicted label by setting a threshold $\tau \in [0, 1]$ to cut off the probability distribution so that inputs with $\mathbb{P} > \tau$ are assigned with positive labels and inputs with $\mathbb{P} < \tau$ are assigned with negative labels.



(a) For predicting **disbelief** labels, linear classifiers achieve best binary-F$_1$ scores near 0.6, and neural-transfer classifiers achieve best binary-F$_1$ scores around 0.8.

(b) For predicting **belief** labels, linear classifiers achieve best binary-F$_1$ scores near 0.5, and neural-transfer classifiers achieve best binary-F$_1$ scores around 0.7.

Figure 5: Precision-recall (for predicting positive labels) curves of six trained classifiers evaluated on the testing set. Three neural transfer-learning based classifiers (BERT, XLNet, and RoBERTa) have similar performance, and outperform two linear classifiers with lexicon-derived features (LIWC+LR and ComLex+LR), which outperform trivial baselines. Isolines for binary-F$_1$ scores are shown.

Before discussing our thresholding strategy (i.e., the choice of τ), we evaluate each classifier on the testing set using precision-recall curves that we obtained by varying τ between 0 and 1. After we choose the threshold τ, we evaluate each classifier on the testing set using unbiasedness (defined later in § 4.3), binary-, macro-, and micro-F$_1$ scores under τ.[9]

Results. The precision-recall curves of all classifiers are shown in Figure 5. Linear classifiers with lexicon-derived features (LIWC+LR and ComLex+LR) outperform trivial baseline methods and achieve their best binary-F$_1$ scores near 0.6 for disbelief (Figure 5a) and 0.5 for belief (Figure 5b). Neural transfer-learning based classifiers (BERT, XLNet and RoBERTa) have the best performance, achieving their best binary-F$_1$ scores around 0.8 for disbelief (Figure 5a) and 0.7 for belief (Figure 5b). The performances of the three neural classifiers are similar, with RoBERTa being slightly better than BERT and XLNet, aligning with the results in (Liu et al. 2019) for generic NLP tasks.

4.3 Thresholding Strategy

In the real world, the thresholding strategy is linked to specific downstream tasks: some common strategies include applying the default $\tau = 0.5$, choosing τ that maximizes

[7] Although longer sequences are truncated to a maximum sequence length, information loss is expected to be rare, considering that commonsense writing styles usually put important (and thus identifiable) content in the beginning of comments (Jiang et al. 2020).

[8] Due to equipment constraints, we are unable to run large models released from these papers.

[9] For binary labels, micro-F$_1$ is equivalent to accuracy.

Table 1: Evaluation results for classification. The chosen thresholds τ, unbiasedness, binary-, macro-, and micro-F_1 scores under τ for all experimented classifiers on the testing set are shown. Chance and linear classifiers can achieve unbiasedness for both disbelief and belief labels but exhibit poor performance. All three neural classifiers can achieve unbiasedness for the disbelief label but only RoBERTa can achieve unbiasedness for the belief label. RoBERTa also has the best F_1 scores.

Classifier	**Disbelief**					**Belief**				
	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1
Chance	0.654	✓	0.354	0.494	0.533	0.814	✓	0.170	0.490	0.691
LIWC+LR	0.415	✓	0.548	0.647	0.675	0.306	✓	0.450	0.666	0.806
ComLex+LR	0.364	✓	0.586	0.683	0.712	0.279	✓	0.371	0.612	0.761
BERT	0.374	✓	0.801	0.840	0.850	0.646	✗	0.620	0.773	0.877
XLNet	0.514	✓	0.798	0.839	0.850	0.593	✗	0.646	0.785	0.877
RoBERTa	0.436	✓	**0.817**	**0.855**	**0.864**	0.451	✓	**0.671**	**0.800**	**0.884**

F_1/accuracy scores, choosing τ under certain precision/recall guarantees, etc.

In our case, however, the application is to use the learned classifier as a proxy for human experts, to measure (dis)belief at scale. Therefore the classifier is expected to make statistically *unbiased* estimations comparing to the underlying label distribution. This means that a desirable τ should equalize error rates between false positives and negatives, so that errors can be balanced out when the classifier is applied onto a large dataset.

Specifically, consider the following confusion matrix:

		Human experts		
		Positive	Negative	
Predictions	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
		TP+FN	FP+TN	N

Consider a tweet expressing (dis)belief as label b, then the underlying prevalence $\mathbb{E}(b)$ in the sample is the number of positive labels (TP+FN) divided by the sample size (N). Using a trained classifier to predict b, the estimated prevalence $\mathbb{E}(\hat{b})$ is then the number of predicted positive labels (TP+FP) divided by the sample size (N). An unbiased classifier should make $\mathbb{E}(b) = \mathbb{E}(\hat{b})$, i.e.,

$$\mathbb{E}(b) = \frac{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}{\mathrm{N}} = \frac{\mathrm{TP}(\tau) + \mathrm{FP}(\tau)}{\mathrm{N}} = \mathbb{E}(\hat{b}), \quad (1)$$

and therefore,

$$\mathrm{FP}(\tau) = \mathrm{FN}(\tau). \quad (2)$$

To verify unbiasedness, we choose a threshold τ using Equation 2 for every classifier from the training set, and then apply the same threshold τ on the testing set and conduct hypothesis tests on Equation 2 again. If Equation 2, as the null hypothesis, is not rejected, the classifier under threshold τ is unbiased. We use the χ^2 test and set the significance level as $p < 0.01$ after Bonferroni correction.

The final evaluation results for all experimented classifiers are shown in Table 1. Chance and linear classifiers, with their simple structure, can easily achieve unbiasedness for both disbelief and belief labels. However, this unbiasedness is moot given their poor performance, as we hypothesize that prevalence will shift in the measurement dataset, i.e., if we apply the Chance classifier under the chosen threshold for measurement, the resulting distribution would be the same as our training data, whose distribution is not representative (as discussed in § 3.1). For the neural classifiers, all three can achieve unbiasedness for the disbelief label but only RoBERTa can achieve unbiasedness for the belief label. In addition, RoBERTa has the best performance evaluated by F_1 scores, therefore we choose it as the classifier to measure (dis)belief at scale.

5 Measurement

As an application of our classifier, we leverage it to measure (dis)belief at scale and explore our proposed research questions. Our measurement study leverages an existing dataset collected by (Jiang and Wilson 2018) that contains 1,672,687 comments collected from Facebook, 113,687 from Twitter, and 828,000 from YouTube written in response to 5,303 fact-checked claims. These claims are drawn from the entire archive of Snopes and PolitiFact's articles between their founding and January 9, 2018.

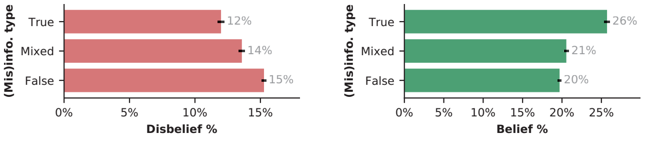
The applicability of our trained classifier on this dataset is suggested by **(a)** the same data collection method, i.e., gathering all comments on social media made in response to seed claims identified from fact check articles; and **(b)** the consistent style of informal English language in social media comments.[10] We preprocess the dataset the same way as our experiments, and then feed the dataset to the RoBERTa-based classifier using our chosen τ as the threshold to predict (dis)belief labels on each comment. This process runs within six hours on a single Titan X Pascal GPU.

5.1 Prevalence of (Dis)belief

(**RQ1**) asks for an estimation of the prevalence of (dis)belief.

This prevalence intuitively varies by the types of (mis)information, therefore we aggregate the veracity of the original claims into three (mis)information types: **(a)** true, if the claims are rated as "true" by Snopes or PolitiFact — these claims contain no misinformation, and their responses were shown to follow distinctive patterns versus others (Jiang and Wilson 2018); **(b)** mixed, if the claims are rated as "mostly true", "half true", or "mixed" — these claims contain some

[10] This, however, does not suggest that the measurement dataset and the training dataset have identical distributions. We are actively working to annotate the measurement dataset for future release.



(a) For **disbelief**, as the veracity of the claims decreases, the prevalence of expressed disbelief increases.

(b) For **belief**, as the veracity of the claims decreases, the prevalence of expressed belief also decreases.

Figure 6: Prevalence of (dis)belief. For true/mixed/false claims on social media, 12%/14%/15% of comments express disbelief and 26%/21%/20% of comments express belief.

misinformation but also some truth; and **(c)** false, if the claims are rated as "mostly false", "false", or "pants on fire!" — these claims contain mostly falsehood.

Next, we aim to estimate the prevalence of (dis)belief in comments in the dataset. However, some of these comments are impacted by a powerful confounding variable: the existence of a fact-check article. To mitigate this, we filter out comments that were posted *after* the corresponding fact-check article was published. Note that, even with this filtering, the remaining comments could still be biased in the claimants distribution (as we discuss in § 6.1).

Finally, we group the remaining comments by the (mis)information type, average their (dis)belief labels (1 if estimated to express (dis)belief and 0 otherwise), and show the results in Figure 6.

We observe that as veracity of claims decrease, disbelief increases while belief decreases. As shown in Figure 6a, we estimate that 12%, 14%, and 15% of comments express disbelief in response to true, mixed, and false claims, respectively; Figure 6b shows that 26%, 21%, and 20% of comments express belief in response to true, mixed, and false claims, respectively. These findings suggests that at least some people commenting on misinformation have the ability to distinguish falsehood, which resonates with the results from existing studies on belief in misinformation (Anderson and Rainie 2017; Mitchell et al. 2014; Nielsen and Graves 2017).

However, the difference in the prevalence of (dis)belief across (mis)information types is relatively small, and for claims that were verified to be true, we estimate that only 26% of comments express belief while 12% express disbelief. One potential explanation for this observation is that the partisan environment drives the public to suspect any claims raised from the opposite ideological group regardless of veracity (Hochschild and Einstein 2015; Guess, Nyhan, and Reifler 2018; Grinberg et al. 2019). Another, though less likely, explanation is that media literacy education equips the public with curiosity to query and doubt all claims, even when the claim is consistent with existing facts (Potter 2018; Hobbs and Jensen 2009). Both explanations are worthy of deeper investigation by future work.

5.2 Effects of Time and Fact-Checks

(**RQ2**) and (**RQ3**) ask for the effects of time and fact-checks. These two questions confound together along the temporal dimension, therefore we investigate them simultaneously. We focus on their effects on false claims, which restricts our analysis to 1,395,293 comments.

Table 2: Regression results for the effects of time and fact-checks. OLS is used to estimate parameters for constant effect ($\hat{\beta}_0$), time effect ($\hat{\beta}_1$), and effect of fact-check ($\hat{\beta}_2$) on 1,395,293 comments in response to false information. There is an extremely slight time effect of falsehood awareness, where disbelief increases 0.001% and belief decreases 0.002% per day after the initial claim. Controlling the time effect, disbelief increases 5% and belief decreases 3.4% after a fact-check.

Parameters	**Disbelief**		**Belief**	
	Estimation	p-value	Estimation	p-value
$\hat{\beta}_0$	$+1.52 \times 10^{-1}$	***	$+1.98 \times 10^{-1}$	***
$\hat{\beta}_1$	$+9.96 \times 10^{-6}$	***	-2.19×10^{-5}	***
$\hat{\beta}_2$	$+5.00 \times 10^{-2}$	***	-3.41×10^{-2}	***
# of samples	$1,395,293$		$1,395,293$	

To investigate (**RQ2**) and (**RQ3**), we formulate the following model: we denote a comment as m, its corresponding claim as C_m, its corresponding fact-check for the claim as F_m, and Δ_{e_1,e_2} as the time difference (unit: days) between event e_1 and event e_2 ($\Delta_{e_1,e_2} > 0$ if e_2 happens after e_1). Then, $\Delta_{C_m,m}$ represents the time delay between a comment and its claim, and $\Delta_{F_m,m}$ represents the time delay between a comment and the fact-check of its claim.

Under these notations, the following model captures the linear effects of time and fact-checks:

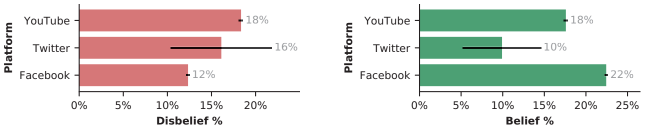
$$\hat{b} = \beta_0 + \underbrace{\beta_1 \cdot \Delta_{C_m,m}}_{\textbf{(RQ2)}} + \underbrace{\beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m})}_{\textbf{(RQ3)}} + \epsilon, \tag{3}$$

where $\hat{b}$ is the underlying prevalence of (dis)belief estimated by the classifier (defined in § 4.3), $\mathbb{I}_+$ is the identity function of positive numbers that returns 1 if the input is positive and 0 otherwise, $\epsilon \sim N(0, \sigma^2)$ is normally distributed noise centered at 0, and β_0, β_1, β_2 are the parameters to be estimated.

This model is similar to the traditional *difference-in-difference* model from causal estimation methods, where the (broadly defined) time variable Δ and the intervention variable $\mathbb{I}$ are regressed jointly to estimate their respected effects (Lechner and others 2011). In our setting, Δ is defined as the time difference between a comment m and its corresponding claim C_m, and $\mathbb{I}$ is a binary variable identified by the time difference between a comment m and its corresponding fact-check F_m.

We use Ordinary Least Square (OLS) to estimate Equation 3 for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$. Here, $\hat{\beta}_0$ represents the constant effects of the underlying initial (dis)belief; $\hat{\beta}_1$ represents the time effect (**RQ2**), i.e., for every unit of $\Delta_{c_m,m}$, (dis)belief is changed by $\hat{\beta}_1$; $\hat{\beta}_2$ represents the effect of fact-checks (**RQ3**), i.e., after fact-checks (the threshold of $\mathbb{I}_+$, $\Delta_{F_m,m} > 0$), (dis)belief is changed by $\hat{\beta}_2$.

As shown in Table 2, there is an extremely slight time effect, where disbelief increases 0.001% and belief decreases



(a) For **disbelief**, Facebook comments express less disbelief than YouTube. However, the difference is not significant for Twitter.

(b) For **belief**, Facebook comments express more belief than YouTube, and YouTube comments express more belief than Twitter.

Figure 7: Platforms difference of expressed (dis)belief. The measured prevalence varies across social media platforms.

0.002% per day after the initial false claims. This effect may be caused by social dynamics, where past comments embed the "wisdom of the crowd" at identifying misinformation, which then impacts future users who engage with the claims (Tschiatschek et al. 2018; Kim et al. 2018). Controlling for the time effect, disbelief increases 5% and belief decreases 3.4% after the publication of a fact-check article, which reinforces existing work on the positive effects of fact-checks (Tambuscio et al. 2015; Hannak et al. 2014; Garrett, Nisbet, and Lynch 2013). Note that although the prevalence of (dis)belief is altered by fact-checks, the mechanism behind such positive effects is still unknown: does the fact-check correct the existing false belief of the same group of users, or does the publication of the fact-check attract a different group of users to comment on the claim with disbelief (therefore altering the overall prevalence)?

5.3 Difference Across Platforms

(**RQ4**) asks for difference in (dis)belief across social media platforms. We process the dataset the same way as § 5.1, except that here we group data by social media platforms instead of misinformation types.

As shown in Figure 7, the prevalence of (dis)belief varies across social media platforms. Figure 7a shows that for disbelief, Facebook comments express less disbelief than YouTube, while the difference is not significant for Twitter. Figure 7b shows that for belief, Facebook comments express more belief than YouTube, whose comments express more belief than Twitter.

Note that this aggregation ignores other confounders, e.g., claim and audience distributions, therefore the result only suggest an overall difference in (dis)belief prevalence across platforms. This reinforces our position (articulated in § 3.1) that analyzing Twitter alone is insufficient to represent the misinformation ecosystem.

6 Discussion

Our study has several limitations that we discuss in this section, together with potential directions for future work.

6.1 Limitations

The dataset we use in our measurement study contains 2,614,374 social media comments, written in response to the entire archive of 5,303 fact-checked claims by Snopes and PolitiFact, a large dataset that is arguably representative to measure the prevalence of (dis)belief. However, the dataset could still be biased in certain respects.

Claimant bias. First, fact-checked claims are, in general, made by high-profile claimants (e.g., political pundits or well-known organizations), therefore excluding claims from the common crowd. There is, to our knowledge, no existing work discussing the relative importance of claims erroneously made (or misinterpreted) by the common crowd in the misinformation ecosystem, therefore we are unable to estimate to what extend this exclusion affects our measurement.

Topical bias. Second, most of the articles from Snopes and PolitiFact are focused on politics or political issues, therefore our measurement is also heavily focused on these topics. Other popular misinformation topics, such as health (Berinsky 2017) or scientific (Farrell, McConnell, and Brulle 2019) misinformation, could be less polarized and thus alter the underlying distributions of (dis)belief.

Proxy validity. The use of comments to understand social interaction is common in social media studies. However, a comment may not reflect the true underlying belief of a person. The Hawthorne effect (McCarney et al. 2007) would suggest that social media users are aware of being observed by the public and thus change their behaviors. Social identity (Stets and Burke 2000) and normative influence theory (Kincaid 2004) would suggest that a comment could be posted just to cater to the preference of a person's ideological group, instead of capturing their true belief. Additionally, the (dis)belief of people who retweet the claim without commenting are not captured in our approach. Therefore, we emphasize that our study measures *expressed* (dis)belief in the misinformation ecosystem, and our results should be interpreted together with existing qualitative and experimental studies (Anderson and Rainie 2017; Nielsen and Graves 2017).

Bots and likewise. Although comments from bot and bot-like (e.g., the Internet Research Agency (IRA)) users are not cleaned in the dataset, recent studies show that bots mostly spread repeated information rather than commenting (Shao et al. 2017), and the IRA had very limited commenting activity comparing to the entire Twitter population (Im et al. 2019; Zannettou et al. 2019). We compared our training dataset verses an IRA account dataset released by Twitter and found no overlap (Gadde and Roth 2018). Therefore, the existence of bots should have minimal effects on our results. Note that the limited commenting activity of IRA does not imply limited *impact*, as a comment can influence subsequent comments. That said, comments under such influence, as long as they are from real users, are intended to be captured in our measurement.

6.2 Future Work

Our work presents initial results for observational studies of expressed (dis)belief in (mis)information, and future work can extend both the modeling and measurement aspects of our work.

Modeling The state-of-the-art pre-trained models for classification in our experiments have been shown to outperform specifically-designed neural architectures for a wide-range of NLP tasks (Devlin et al. 2019; Yang et al. 2019; Liu et al. 2019), and achieved reasonable performance for our task as evaluated in § 4.2. However, there is still potential modeling space to improve classification performance, and we hope the release of our annotated dataset can benefit future researchers for this task.

Measurement Our measurement is first focused on estimating a *coarse average* of (dis)belief prevalence, and then on potential effects of time, fact-checks, and platforms. Future work can apply our released classifier to study more fine-grained research questions, as long as more features are observed in future dataset. In addition to the followup questions we raised in § 5, there are other potential directions along this line, e.g., a recent study showed that conservatives and senior citizens are more vulnerable to *spread* misinformation (Guess, Nagler, and Tucker 2019), but are they also more vulnerable to *believe* it, and if so, to what extend? Are there geographic or longitudinal differences in the distribution of (dis)belief?

6.3 Conclusion

In this paper, we proposed and developed an observational approach to understand expressed (dis)belief in (mis)information by leveraging comments as a proxy. We applied our trained classifier to explore the research questions of the prevalence of (dis)belief, the effects of time and fact-checks, and differences across social media platforms.

Our measurement delivered some optimistic results, e.g., increased disbelief and decreased belief as information veracity decrease, (albeit slightly) increased disbelief and decreased belief for false claims over time, a positive effect of fact-checks. However, these results do not undermine the fundamentally concerning consequences of misinformation, especially since we also found some pessimistic results, e.g., considerable suspicion of truthful claims.

Despite several notable limitations, we hope this work will be a helpful addition to the literature that complements existing qualitative and experimental studies of (dis)belief and (mis)information.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and Ming Min for invaluable discussion on this work. This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2).

Anderson, J., and Rainie, L. 2017. The future of truth and misinformation online. *Pew Research Center*.

Berinsky, A. J. 2017. Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science* 47(2).

Burfoot, C., and Baldwin, T. 2009. Automatic satire detection: Are you having a laugh? In *Proc. of ACL*.

Ciampaglia, G. L.; Mantzarlis, A.; Maus, G.; and Menczer, F. 2018. Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine* 39(1).

Dai, Z.; Yang, Z.; Yang, Y.; Cohen, W. W.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proc. of ACL*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Epstein, R., and Robertson, R. E. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *PNAS* 112(33).

Farajtabar, M.; Yang, J.; Ye, X.; Xu, H.; Trivedi, R.; Khalil, E.; Li, S.; Song, L.; and Zha, H. 2017. Fake news mitigation via point process based intervention. In *Proc. of ICML*.

Farías, D. I. H.; Patti, V.; and Rosso, P. 2016. Irony detection in twitter: The role of affective content. *ACM ToIT* 16(3).

Farrell, J.; McConnell, K.; and Brulle, R. 2019. Evidence-based strategies to combat scientific misinformation. *Nature Climate Change*.

Fletcher, R.; Cornia, A.; Graves, L.; and Nielsen, R. K. 2018. Measuring the reach of ?fake news? and online disinformation in europe. *Reuters institute factsheet*.

Gadde, V., and Roth, Y. 2018. Enabling further research of information operations on twitter. *Twitter Blog* 17.

Garrett, R. K.; Nisbet, E. C.; and Lynch, E. K. 2013. Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naïve theory. *Journal of Communication* 63(4).

Gentzkow, M.; Kelly, B.; and Taddy, M. 2019. Text as data. *Journal of Economic Literature* 57(3).

González-Ibánez, R.; Muresan, S.; and Wacholder, N. 2011. Identifying sarcasm in twitter: a closer look. In *Proc. of HLT*.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on twitter during the 2016 us presidential election. *Science* 363(6425).

Guess, A.; Nagler, J.; and Tucker, J. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances* 5(1).

Guess, A.; Nyhan, B.; and Reifler, J. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*.

Hannak, A.; Margolin, D.; Keegan, B.; and Weber, I. 2014. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *Proc. of ICWSM*.

Hasan, K. S., and Ng, V. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proc. of IJCNLP*.

Hobbs, R., and Jensen, A. 2009. The past, present, and future of media literacy education. *Journal of media literacy education* 1(1).

Hochschild, J. L., and Einstein, K. L. 2015. *Do facts matter?: Information and misinformation in American politics*, volume 13. University of Oklahoma Press.

Hu, D.; Jiang, S.; Robertson, R. E.; and Wilson, C. 2019. Auditing the partisanship of google search snippets. In *Proc. of WWW*.

Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; and Gilbert, E. 2019. Still out there: Modeling and identifying russian troll accounts on twitter. *arXiv*.

Jang, S. M., and Kim, J. K. 2018. Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior* 80.

Jiang, S., and Wilson, C. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *PACM on HCI* 2(CSCW).

Jiang, S.; Baumgartner, S.; Ittycheriah, A.; and Yu, C. 2020. Factoring fact-checks: Structured information extraction from fact-checking articles. In *Proc. of WWW*.

Jiang, S.; Robertson, R. E.; and Wilson, C. 2019. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proc. of ICWSM*.

Jiang, S.; Robertson, R. E.; and Wilson, C. 2020. Reasoning about political bias in content moderation. In *Proc. of AAAI*.

Joseph, K.; Friedland, L.; Hobbs, W.; Lazer, D.; and Tsur, O. 2017. Constance: Modeling annotation contexts to improve stance classification. In *Proc. of EMNLP*.

Kaur, K.; Nair, S.; Kwok, Y.; Kajimoto, M.; Chua, Y. T.; Labiste, M.; Soon, C.; Jo, H.; Lin, L.; Le, T. T.; et al. 2018. Information disorder in asia and the pacific: Overview of misinformation ecosystem in australia, india, indonesia, japan, the philippines, singapore, south korea, taiwan, and vietnam. *SSRN*.

Kim, J.; Tabibian, B.; Oh, A.; Schölkopf, B.; and Gomez-Rodriguez, M. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proc. of WSDM*.

Kincaid, D. L. 2004. From innovation to social norm: Bounded normative influence. *Journal of health communication* 9(S1).

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In *Proc. of AAAI*.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science* 359(6380).

Lechner, M., et al. 2011. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics* 4(3).

Levendusky, M. S. 2013. Why do partisan media polarize viewers? *American Journal of Political Science* 57(3).

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*.

Liu, Y.; Kliman-Silver, C.; and Mislove, A. 2014. The tweets they are a-changin': Evolution of twitter users and behavior. In *Proc. of ICWSM*.

McCarney, R.; Warner, J.; Iliffe, S.; Van Haselen, R.; Griffin, M.; and Fisher, P. 2007. The hawthorne effect: a randomised, controlled trial. *BMC Medical Research Methodology* 7(1).

Mitchell, A.; Gottfried, J.; Kiley, J.; and Matsa, K. E. 2014. Political polarization & media habits. *Pew Research Center* 21.

Morgan, S. 2018. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy* 3(1).

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proc. of ICWSM*.

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2).

Nielsen, R. K., and Graves, L. 2017. "news you don't believe": Audience perspectives on fake news. *Reuters Institute*.

Pan, S. J., and Yang, Q. 2009. A survey on transfer learning. *IEEE TKDE* 22(10).

Potter, W. J. 2018. *Media literacy*. Sage Publications.

Poynter. 2020. Verified signatories of the ifcn code of principles.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proc. of ACL*.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing partisan audience bias within google search. *PACM on HCI* 2(CSCW).

Robertson, R. E.; Jiang, S.; Lazer, D.; and Wilson, C. 2019. Auditing autocomplete: Suggestion networks and recursive algorithm interrogation. In *Proc. of WebSci*.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of misinformation by social bots. *arXiv*.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1).

Stets, J. E., and Burke, P. J. 2000. Identity theory and social identity theory. *Social psychology quarterly*.

Street, C. T., and Ward, K. W. 2012. Improving validity and reliability in longitudinal case study timelines. *European Journal of Information Systems* 21(2).

Tambuscio, M.; Ruffo, G.; Flammini, A.; and Menczer, F. 2015. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proc. of WWW*.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1).

Tschiatschek, S.; Singla, A.; Gomez Rodriguez, M.; Merchant, A.; and Krause, A. 2018. Fake news detection in social networks via crowd signals. In *Companion Proc. of WWW*.

Twitter. 2020. About different types of tweets.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Ward, A.; Ross, L.; Reed, E.; Turiel, E.; and Brown, T. 1997. Naïve realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*.

Wasserman, H., and Madrid-Morales, D. 2019. An exploratory study of ?fake news? and media trust in kenya, nigeria and south africa. *African Journalism Studies* 40(1).

Xing, Z.; Pei, J.; and Keogh, E. 2010. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter* 12(1).

Yang, D.; Lavie, A.; Dyer, C.; and Hovy, E. 2015. Humor recognition and humor anchor extraction. In *Proc. of EMNLP*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of NeurIPS*.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtelris, N.; Leontiadis, I.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proc. of IMC*.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019. Who let the trolls out? towards understanding state-sponsored trolls. In *Proc. of WebSci*.

Zhou, X.; Zafarani, R.; Shu, K.; and Liu, H. 2019. Fake news: Fundamental theories, detection strategies and challenges. In *Proc. of WSDM*.

Zhou, X.; Wan, X.; and Xiao, J. 2016. Attention-based lstm network for cross-lingual sentiment classification. In *Proc. of EMNLP*.