# Can You Verifi This? Studying Uncertainty and Decision-Making About Misinformation using Visual Analytics

**Alireza Karduni[1*], Ryan Wesslen[1], Sashank Santhanam[1], Isaac Cho[1],**

**Svitlana Volkova[2], Dustin Arendt[2], Samira Shaikh[1], Wenwen Dou[1]**

[1]Department of Computer Science, UNC-Charlotte

[2]Pacific Northwest National Laboratory

[*]akarduni@uncc.edu

## Abstract

We describe a novel study of decision-making processes around misinformation on social media. Using a custom-built visual analytic system, we presented users with news content from social media accounts from a variety of news outlets, including outlets engaged in distributing misinformation. We conducted controlled experiments to study decision-making regarding the veracity of these news outlets and tested the role of confirmation bias (the tendency to ignore contradicting information) and uncertainty of information on human decision-making processes. Our findings reveal that the presence of conflicting information, presented to users in the form of cues, impacts the ability to judge the veracity of news in systematic ways. We also find that even instructing participants to explicitly disconfirm given hypotheses does not significantly impact their decision-making regarding misinformation when compared to a control condition. Our findings have the potential to inform the design of visual analytics systems so that they may be used to mitigate the effects of cognitive biases and stymie the spread of misinformation on social media.

## Introduction

The spread of misinformation on social media is a phenomena with global consequences, one that, according to the World Economic Forum, poses significant risks to democratic societies (Howell and others 2013). The online media ecosystem is now a place where false or misleading content resides on an equal footing with verified and trustworthy information (Kott, Alberts, and Wang 2015). In response, social media platforms are becoming "content referees," faced with the difficult task of identifying misinformation internally or even seeking users' evaluations on news credibility.[1] On the one hand, the news we consume is either wittingly or unwittingly self-curated, even self-reinforced (Tsang and Larson 2016). On the other hand, due to the explosive abundance of media sources and the resulting information overload, we often need to rely on heuristics and social cues to make decisions about the credibility of information (Mele et al. 2017; Shao et al. 2017). One such decision-making heuristic is confirmation bias, which has been implicated in the selective exposure to and spread of misinformation (Allan 2017). This cognitive bias can manifest itself on social media as individuals tend to select claims and consume news that reflect their preconceived beliefs about the world, while ignoring dissenting information (Mele et al. 2017).

While propaganda and misinformation campaigns are not a new phenomenon (Soll 2017), the ubiquity and virality of the internet has lent urgency to the need for understanding how individuals make decisions about the news they consume and how technology can aid in combating this problem (Shu et al. 2017). Visual analytic systems that present coordinated multiple views and rich heterogeneous data have been demonstrably useful in supporting human decision-making in a variety of tasks such as textual event detection, geographic decision support, malware analysis, and financial analytics (Wagner et al. 2015; Wanner et al. 2014). **Our goal is to understand how visual analytics systems can be used to support decision-making around misinformation and how uncertainty and confirmation bias affect decision-making within a visual analytics environment**.

In this work, we seek to answer the following overarching research questions: *What are the important factors that contribute to the investigation of misinformation? How to facilitate decision-making around misinformation by presenting the factors in a visual analytics system? What is the role of confirmation bias and uncertainty in such decision-making processes?* To this aim, we first leveraged prior work on categorizing misinformation on social media (specifically Twitter) (Volkova et al. 2017) and identified the dimensions that can distinguish misinformation from legitimate news. We then developed a visual analytic system, Verifi, to incorporate these dimensions into interactive visual representations. Next, we conducted a controlled experiment in which participants were asked to investigate news media accounts using Verifi. Through quantitative and qualitative analysis of the experiment results, we studied the factors in decision-making around misinformation. More specifically, we investigated how **uncertainty, conflicting signals manifested in the presented data dimensions**, affect users' ability to identify misinformation in different experiment conditions. Our work is thus uniquely situated at the intersection of the psychology of decision-making, cognitive biases, and the impact of socio-technical systems, namely visual analytic



[1]https://www.wsj.com/articles/facebook-to-rank-news-sources-by-quality-to-battle-misinformation-1516394184

systems, that aid in such decision-making.

Our work makes the following important contributions:

- *A new visual analytic system*: We designed and developed Verifi[2], a new visual analytic system that incorporates dimensions critical to characterizing and distinguishing misinformation from legitimate news. Verifi enables individuals to make informed decisions about the veracity of news accounts.
- *Experiment design to study decision-making on misinformation*: We conducted an experiment using Verifi to study how people assess the veracity of the news media accounts on Twitter and what role confirmation bias plays in this process. To our knowledge, our work is the first experimental study on the determinants of decision-making in the presence of misinformation in visual analytics.

As part of our controlled experiment, we provided cues to the participants so that they would interact with data for the various news accounts along various dimensions (e.g.,tweet content, social network). Our results revealed that conflicting information along such cues (e.g., connectivity in social network) significantly impacts the users' performance in identifying misinformation.

## Related Work

We discuss two distinct lines of past work that are relevant to our research. First, we explore cognitive biases, and specifically the study of confirmation bias in the context of visual analytics. Second, we introduce prior work on characterizing and visualizing misinformation in online content.

**Confirmation bias:** Humans exhibit a tendency to treat evidence in a biased manner during their decision-making process in order to protect their beliefs or pre-conceived hypothesis (Jonas et al. 2001), even in situations where they have no personal interest or material stake (Nickerson 1998). Research has shown that this tendency, known as confirmation bias, can cause inferential error with regards to human reasoning (Evans 1989). Confirmation bias is the tendency to privilege information that confirms one's hypotheses over information that disconfirms the hypotheses. Classic laboratory experiments to study confirmation bias typically present participants with a hypothesis and evidence that either confirms or disconfirms their hypothesis, and may include cues that cause uncertainty in interpretation of that given evidence. Our research is firmly grounded in these experimental studies of confirmation biases. We adapt classic psychology experimental design, where pieces of evidence or *cues* are provided to subjects used to confirm or disconfirm a given hypothesis (Wason 1960; Nickerson 1998).

**Visualization and Cognitive Biases:** Given the pervasive effects of confirmation bias and cognitive biases in general on human decision-making, scholars studying visual analytic systems have initiated research on this important problem. (Wall et al. 2017) categorized four perspectives to build a framework of all cognitive biases in visual analytics. (Cho et al. 2017) presented a user study and identified an approach to measure anchoring bias in visual analytics by priming users to visual and numerical anchors. They demonstrated that cognitive biases, specifically anchoring bias, affect decision-making in visual analytic systems, consistent with prior research in psychology. However, no research to date has examined the effects of confirmation bias and uncertainty in the context of distinguishing information from misinformation using visual analytic systems - we seek to fill this important gap. Next, we discuss what we mean by misinformation in the context of our work.

**Characterizing Misinformation:** Misinformation can be described as information that has the camouflage of traditional news media but lacks the associated rigorous editorial processes (Mele et al. 2017). Prior research in journalism and communication has demonstrated that news outlets may slant their news coverage based on different topics (Entman 2007). In addition, (Allcott and Gentzkow 2017) show that the frequency of sharing and distribution of fake news can heavily favor different individuals. In our work, we use the term fake news to encompass misinformation including ideologically slanted news, disinformation, propaganda, hoaxes, rumors, conspiracy theories, clickbait and fabricated content, and even satire. We chose to use "fake news" as an easily accessible term that can be presented to the users as a label for misinformation and we use the term "real news" as its antithesis to characterize legitimate information.

Several systems have been introduced to (semi-) automatically detect misinformation, disinformation, or propaganda in Twitter, including FactWatcher (Hassan et al. 2014), TwitterTrails (Metaxas, Finn, and Mustafaraj 2015), RumorLens (Resnick et al. 2014), and Hoaxy (Shao et al. 2016). These systems allow users to explore and monitor detected misinformation via interactive dashboards. They focus on identifying misinformation and the dashboards are designed to present analysis results from the proposed models. Instead, Verifi aims to provide an overview of dimensions that distinguish real vs. fake news accounts for a general audience.

Our work is thus situated at the intersection of these research areas and focuses on studying users' decision making about misinformation in the context of visual analytics.

## Verifi: A Visual Analytic System for Investigating Misinformation

Verifi is a visual analytic system that presents multiple dimensions related to misinformation on Twitter. Our design process is informed by both prior research in distinguishing real and fake news as well as our analysis based on the data selected for our study to identify meaningful features.

A major inspiration for Verifi's design is based on the findings of (Volkova et al. 2017), who created a predictive model to distinguish between four types of fake news accounts. They find that attributes such as *social network interactions* (e.g., mention or retweet network), *linguistic features*, and *temporal trends* are the most informative factors for predicting the veracity of Twitter news accounts. Our design of Verifi is inspired by these findings: (i) we included

[2]http://verifi.herokuapp.com; open source data and code provided at https://github.com/wesslen/verifi-icwsm-2018

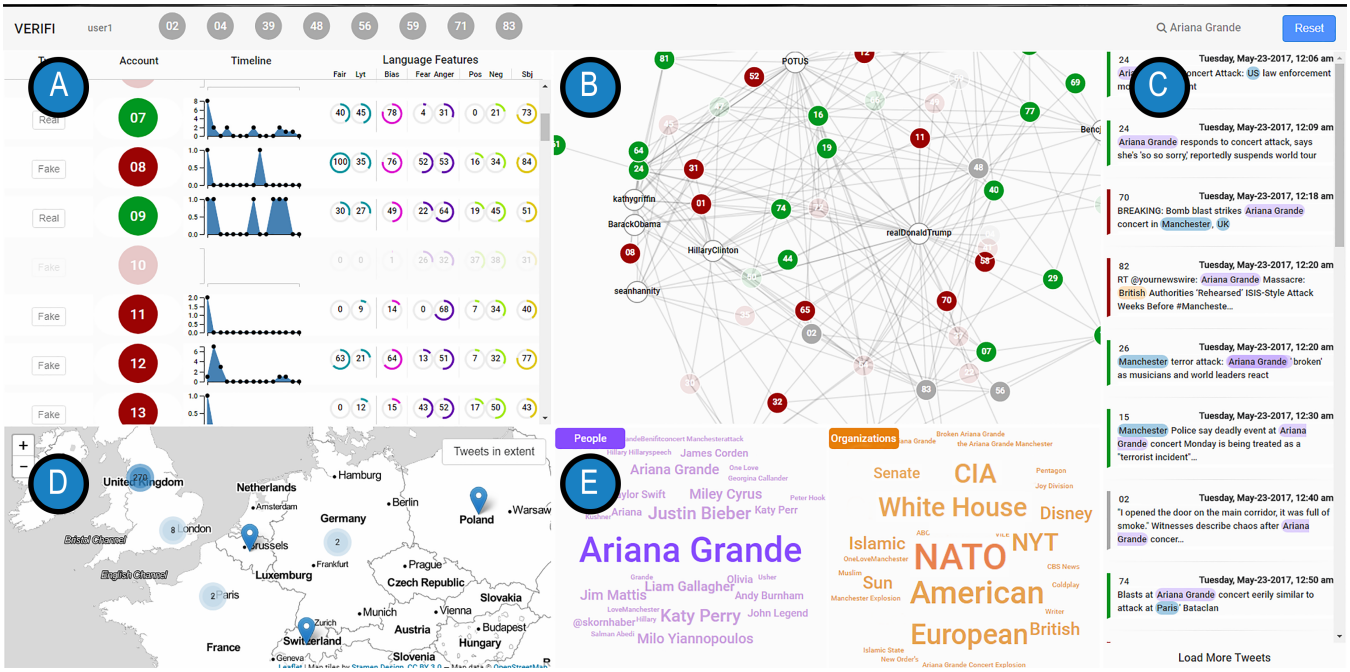


Figure 1: The Verifi interface: Account View (A), Social Network View (B), Tweet Panel (C), Map View (D), and Entity Word Cloud (E). The interface can be accessed at Verifi.Herokuapp.com.

a *social network view* that shows a visualization of account mentions (which includes retweets) as a primary view to allow users to investigate relationships between accounts; (ii) we developed an accounts view with *account-level temporal (daily) trends* as well as the most predictive linguistic features to facilitate users' account-level investigation into the rhetoric and timing of each account's tweets; and (iii) to choose the most effective *linguistic features*, we created a model to predict which linguistic features most accurately can predict the veracity of different accounts.

In addition to three different analytical cues inspired by Volkova et al. and our predictive model, we included visualizations and data filtering functions to allow participants to qualitatively examine and compare accounts. Based on existing research conducted on the ways news can be slanted and the diffusion of misinformation (Adams 1986; Allcott and Gentzkow 2017; Entman 2004; 2007), we included visual representation of three types of extracted *entities (places, people, and organizations)* to enable exploration through filtering.

## Dataset

To create our dataset, we started with a list of 147 Twitter accounts annotated as propaganda, hoax, clickbait, or satire by (Volkova et al. 2017) based on public sources. We then augmented this list with 31 mainstream news accounts (Starbird 2017) that are considered trustworthy by independent third-parties.[3] We collected 103,248 tweets posted by these 178 accounts along with account metadata from May 23, 2017 to June 6, 2017 using the Twitter public API.[4]

| Type | Real | Propaganda | Clickbait | Hoax | Satire |
|---|---|---|---|---|---|
| Account | 31 | 30 | 18 | 2 | 2 |

Table 1: Distribution of types of news outlets

We then filtered the 178 accounts using the following criteria indicating that the account is relatively less active: (i) low tweet activity during our data collection period; (ii) recent account creation date; and (iii) low friends to follower ratio. In addition to these three criteria, we asked two trained annotators to perform a qualitative assessment of the tweets published by the accounts and exclude extreme accounts (e.g., highly satirical) or non-English accounts. After these exclusions, we had a total of 82 accounts, distributed along the categories shown in Table 1.

## Data processing and analysis

To analyze our tweet data, we extracted various linguistic features, named entities, and social network structures. The role of the computational analysis in our approach is to support hypothesis testing based on social data driven by social science theories (Wallach 2018).

**Language features:** Language features can characterize the style, emotion, and sentiment of news media posts. Informed by prior research that identified multiple language features for distinguishing real versus fake news (Volkova

[3]https://tinyurl.com/yctvve9h and https://tinyurl.com/k3z9w2b

[4]The Verifi interface relies on a public Twitter feed collected by the University of North Carolina Charlotte.

| Source | Features | Example |
|---|---|---|
| Bias Language Lexicon-driven | 6 | Bias, Factives, Implicatives, Hedges, Assertives, Reports |
| Moral Foundation Lexicon-driven | 11 | Fairness, Loyalty, Authority, Care |
| Subjectivity Lexicon-driven | 8 | Strong Subjective, Strong Negative Subjective, Weak Neutral Subjective |
| Sentiment Model-driven | 3 | Positive, Negative, Neutral |
| Emotions Model-driven | 5 | Anger, Disgust, Fear, Joy, Sadness, Surprise |

Table 2: 34 candidate language features from five sources.

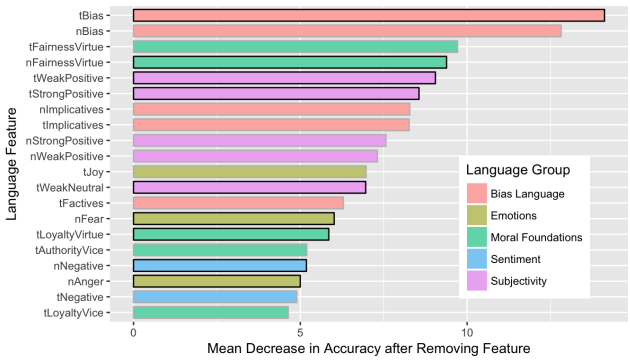


Figure 2: Top 20 most predictive language features of Fake and Real news outlets as measured by each feature's average effect on Accuracy. 't' prefix indicates the feature is normalized by the account's tweet count and 'n' indicates normalization by the account's word count (summed across all tweets). Features with borders are included in Verifi.

et al. 2017), we consider five language features, including *bias language* (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013), *subjectivity* (Wilson, Wiebe, and Hoffmann 2005), *emotion* (Volkova et al. 2015), *sentiment* (Liu, Hu, and Cheng 2005), and *moral foundations* (Graham, Haidt, and Nosek 2009; Haidt and Graham 2007). For example, *moral foundations* is a dictionary of words categorized along eleven dimensions, including care, fairness, and loyalty. Table 2 provides an overview of the features we used to characterize the language of the tweets, with each feature containing multiple dimensions.

In total, we test 68 different dimensions (i.e., 34 different language feature dimensions and each with two different normalization methods – either by number of tweets or number of words) using a supervised machine-learning algorithm (Random Forest) with a 70/30 training/validation split. We eliminated highly correlated (redundant) features (see supplemental materials). Figure 2 provides the ranking of the top 20 predictive language features.[5] Using this ranking, we decided to include eight language features within Verifi: *Bias, Fairness (as a virtue), Loyalty (as a virtue), Negative sentiment, Positive sentiment, Fear, and Subjectivity* to assist users in distinguishing fake and real news.[6]

**Entity Extraction and Geocoding:** Verifi includes a word cloud to display the top mentioned entities and enable the comparison of how different media outlets talk about entities of interest. We extract people, organization, and location entities from the tweets.

**Social Network Construction:** To present the interactions between the accounts on Twitter, we construct an undirected social network. Edges are mentions or retweets between accounts. Nodes represent Twitter news accounts (82 nodes) as well as the top ten most frequently mentioned Twitter accounts by our selected accounts.

## The Verifi User Interface

The Verifi user interface is developed using D3.js, Leaflet, and Node.js. The interface consists of six fully coordinated views that allow users to explore and make decisions regarding the veracity of news accounts (Figure 1).

**The Accounts View** (Figure 1A) provides account-level information including tweet timeline and language features. The circular button for each account is color coded to denote whether the account is considered real (green) or fake (red). The accounts colored in gray are the ones participants are tasked to investigate in our experiment. The timeline shows the number of tweets per day. The array of donut charts shows the eight selected language features (scaled from 0-100) that characterize the linguistic content. For example, a score of 100 for fairness means that an account exhibits the highest amount of fairness in its tweets compared to the other accounts. Users can sort the accounts based on any language feature. The Account View provides an overview of real and fake accounts and enables analysis based on language features and temporal trends.

**The Social Network View** (Figure 1B) presents connections among news accounts (nodes) based on mentions and retweets (edges). The color coding of the nodes is consistent with the Accounts View (i.e., green for real, red for fake, gray for unknown). To increase the connectivity of the news accounts, we included ten additional Twitter accounts. These ten accounts (colored white) are the top-ranked Twitter accounts by mention from the 82 news accounts over the two week period. The Social Network View allows users to understand how a specific account is connected to fake or real news accounts on the social network.

**Entity Views:** The people and organization word clouds (Figure 1E) present an overview of the most frequently mentioned people and organization entities. The word clouds support the filtering of tweets mentioning certain entities of interest, thus enabling comparison across accounts. For example, by clicking on the word "American," accounts that mention this entity would be highlighted in both the Accounts View and the Social Network View. In addition, tweets mentioning "American" will appear in the Tweet

[5]This model had a 100% validation accuracy (24 out of 24) on the 30% validation dataset.

[6]We averaged Strong-Weak subjectivity measures into one single measure.

| Mask Name | Description |
|---|---|
| @XYZ | A news division of a major broadcasting company |
| @GothamPost | An American newspaper with worldwide influence and readership |
| @MOMENT | An American weekly news magazine |
| @Williams | An international news agency |
| @ThirtyPrevent | A financial blog with aggregated news and editorial opinions |
| @ViralDataInc | An anti right-wing news blog and aggregator |
| @NationalFist | An alternative media magazine and online news aggregator |
| @BYZBrief | Anti corporate propaganda outlet with exclusive content and interviews |

Table 3: Eight accounts with masked account names. Background colors indicate real (green) and fake (red).

Panel View.

**The Map View** provides a summary of the location entities (Figure 1D). When zooming in and out, the color and count of the cluster updates to show the tweets in each region. Users can click on clusters and read associated tweets. Users can also filter data based on a geographic boundary.

**The Tweet Panel View** (Figure 1C) provides drill-down capability to the tweet level. Users can use filtering to inspect aggregate patterns found in other views. Within the tweet content, detected entities are highlighted to assist users in finding information in text. This view is similar to how Twitter users typically consume tweets on mobile devices.

## Experiment Design

We designed a user experiment to study how people make decisions regarding misinformation and the veracity of new accounts on Twitter with the help of the Verifi system.

### Research Questions

Situated in the context of decision-making with visual analytics, we organized our research focus on the following research questions:

**RQ1:** Would individuals make decisions differently about the veracity of news media sources, when *explicitly asked to confirm or disconfirm* a given hypothesis?

**RQ2:** How does uncertainty (conflicting information) of cues affect performance on identifying accounts that post misinformation?

### Experiment Stimuli

After developing the Verifi interface, we loaded data from all 82 accounts (Table 1) into the system. To minimize the effect of preconceived notions, all news outlet names were anonymized by assigning them integer identifiers. Given the in-lab nature of the user studies and time limitations, we selected eight accounts that participants would investigate and would label as either real or fake based on their own judgments. The accounts were chosen to cover a range of different cues and degrees of uncertainty. We based our selection of experiment stimuli on classic studies in confirmation bias (Wason 1960; Rajsic, Wilson, and Pratt 2015).

Due to institutional concerns, we have masked the names of those accounts while preserving the nature of their naming. The eight selected accounts (4 real and 4 fake) with their masked names and description are shown in Table 3. The source of the description is Wikipedia and identifying information was removed to anonymize the accounts. Our goal in selecting the experiment stimuli was to enable participants to make decisions about a wide range of content with the aid of varied, sometimes conflicting, cues.

### Experiment Tasks

To test the effect of confirmation bias, we designed an experiment with three experimental conditions: Confirm, Disconfirm, and Control. In the Confirm condition, participants were given a set of six cues about the grayed out accounts (i.e.,the eight selected accounts shown in Table 3) and were explicitly asked to *confirm* a given hypothesis that all grayed-out accounts were fake accounts. Similarly, in the Disconfirm condition, participants were explicitly asked to *disconfirm* the given hypothesis that all gray accounts were fake. Our third experiment condition was the Control, where the participants were simply asked to judge the veracity of the accounts; they were given neither the initial hypothesis nor the set of six cues. Following classic psychology studies in confirmation bias (Nickerson 1998) where the information presented to the participants has inherent uncertainty, we added the element of uncertainty to the cues. We provided six cues (Q1-Q6) to the participant, of which three cues pointed to the account being real and three cues pointing to the account being fake. Each cue corresponds to a view in the Verifi interface.

The decisions that participants needed to make for the gray accounts involved answering (True/False/Did Not Investigate) for each of the six statements listed below. Each statement is the same as the cue presented to the participant in the confirm and disconfirm condition; the purpose of the statements is to gather information on which cues the participants relied on when making decisions for a certain account.

**Q1** This account is predominantly connected to real news accounts in the **social network** graph. This characteristic is typically associated with known real news accounts.

**Q2** The average **rate of tweeting** from this account is relatively low (less than 70 tweets per day).

**Q3** On the language measures, this account tends to show a higher ranking in **bias measure and fairness measure**. This characteristic is typically associated with known real news accounts.

**Q4** This account tends to focus on a subset of polarizing **entities** (people, organizations, locations) such as Barack Obama or Muslims as compared to focusing on a diverse range of entities.

**Q5** On the language measures, this account tends to show a low ranking in **fear and negative** language measures.

**Q6** The tweets from this account contain **opinionated language**. This characteristic is typically associated with

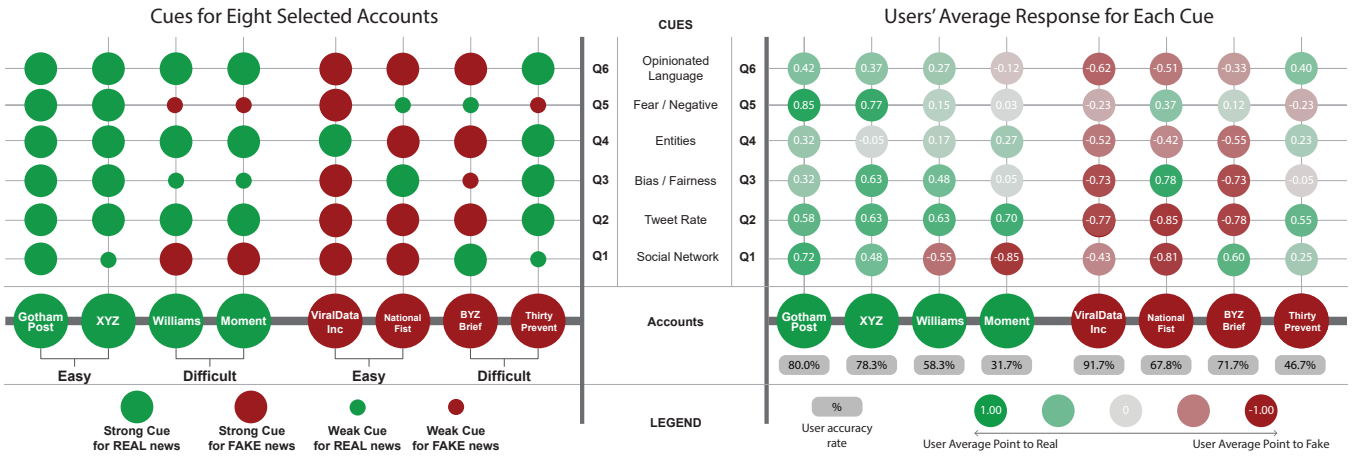


Figure 3: Available cues for selected accounts (column) and users' response regarding the importance of these cues (row, Q1-Q6). Left: Shows each of the eight selected accounts as well as the cues available for each of them. Right: Shows average of importance for each cue per account based on participants' responses. Values in gray circles below each account name show average accuracy for predicting that account correctly. The left figure is purely based on the (conflicting) information presented in the cues and is independent from user responses. The right figure based on the user responses on the importance of each cue coincides with the information in the left table.

known fake news accounts.

These statements and the cues given to the participants at the beginning of the experiment (along with the hypothesis) are the same. Based on our data collection and analysis, statements Q1, Q3, and Q5 point to an account while being a real account and while the rest of the statements (Q2, Q4, and Q6) point to the account being fake. For certain statements, we explicitly included information characterizing whether the cue pointed to the account being real or fake (as shown in Q1, Q3, and Q6). The presentation of these cues were deliberately chosen to add to the uncertainty of information presented to the users. In addition to asking participants about their decision-making process on each statement listed above, we also asked the users to rate the importance of each view in the Verifi interface in making those decisions (the Accounts View, Social Network View, Tweet Panel View, and Entity View) on a scale of 1 to 7. Additionally, we asked participants to indicate the confidence of their decision on a scale of 1 to 7 for each account, as well as an optional, free-form response section where participants could provide any additional information as a part of their analysis. All these questions were part of a pop-up form that was displayed when the participants clicked the "Choice" button shown alongside the account number in the Accounts View (Figure 1A). The responses to this form were captured in a database upon submitting the form during the task.

The information regarding each gray account and its cues is summarized in Figure 3 (Left). For simplicity of presentation, green circles indicate a cue pointing towards account being real, red circles indicate a cue pointing towards account being fake. The overview of how they score on cues demonstrate how the accounts exhibit different levels of difficulty for decision-making. For example, all evidence pointed to the *@GothamPost* account being real, which means that ideally, upon investigation, a participant would answer True for Q1, Q3 and Q5 and False for Q2, Q4 and Q6 when making their decision for that account. However, other real accounts chosen for investigation had more uncertainty in the cues. Notably, the *@MOMENT* account was chosen as one of the difficult accounts since it had a misleading social network cue (Q1) in that it had only one connection to a fake news account. For the fake news accounts chosen, *@ViralDataInc* had all evidence pointing towards the account being fake (except Q4, which means that the tweets from this account covered a diverse range of entities). This would make *@ViralDataInc* easier to judge as fake than, for instance, *@ThirtyPrevent*, which exhibits many more misleading cues.

## Experiment Procedure and Participants

We recruited participants via in class recruitment, email to listservs, and the psychology research pool at our institution. Once signed up, participants came to the lab for a duration of one hour. After the informed consent procedure, participants viewed two training videos designed for this experiment. The first video introduced the interface and explained the different views. The second video provided a task example to determine the veracity of a sample account not used in the study. Both videos were identical across all conditions. After this training, participants completed a pre-questionnaire consisting of questions related to their demographics (age, gender, education), familiarity with visual analytics, social media, and Big-5 personality questions (John, Donahue, and Kentle 1991). The participants were then assigned the task and asked to complete the task in 30 minutes. After completing their task, participants completed a post-test questionnaire which included six vignettes to assess participant's propensity to confirmation bias in general (Nickerson 1998).

Sixty participants completed the study, evenly split into three treatment groups. Participant ages were between 18 and 41 (mean=24.7). The gender distribution was 45% male and 55% female. A majority of the participants were undergraduates (65%), followed by Master's (16.7%), Ph.D. (8.3%), and others (10%). The distributions of the participants between computing (48.3%) and non-computing majors (51.7%) was relatively even.

## Data Analysis Methods

In this section, we introduce the analysis methods applied to our experiment data to answer the two research questions.

**To address RQ1**, namely, "are there significant differences in the way participants interact with the data and their resulting judgments based on the experiment condition?", we use one-way analysis of variance (ANOVA) for testing and post-hoc Tukey's honest significant difference (HSD) test to determine significance ($\alpha$=0.05). Our experiment design is a between-subjects design with one level: the experimental condition.

**To address RQ2** regarding the effects of uncertainty, we used two logistic regressions to explore the effects of uncertainty (in cues, accounts, confidence, and treatment groups) had on users' decision-making. Each regression included a different dependent variable: users' accuracy (1 = correct decision, 0 = incorrect decision) and fake determination (1 = fake prediction, 0 = real prediction). This analysis allows us to determine which factors were most important and aligned with our expectation in terms of direction. For example, as mentioned in the Experiment Stimuli section, cues Q1, Q3, and Q5 were selected to point to real accounts, suggesting a negative relationship with fake prediction (or less than 1 log odds ratios). Alternatively, cues Q2, Q4, and Q6 were selected to point to fake accounts (i.e., positive relationship or greater than 1 log odds ratios). In addition, we can also identify which cue was most important in decision-making as the one with the largest (in absolute magnitude) coefficient. In addition to the cues, we also include dummy variables for the account-level (using @XYZ as the reference level) as well as include users' confidence level and treatment group (Control group is the reference level) to understand if these factors played an additional role in the users' decisions.

## Analyses Results

In this section, we describe our findings and results. The detailed discussion about the implications of these findings is in the Discussion section.

### RQ1: Testing the Effects of Confirmation Bias

Table 4 shows the user accuracy rate and fake prediction rate across all three experiment conditions. We found no significant differences between the experimental conditions, on a diverse range of factors. Participants in all three conditions did not differ on the number of accounts labeled as fake and the number of accounts labeled as real ($p$>0.05 for both). We tested the accuracy rate and found no significant difference in the rate of accuracy across experimental conditions ($p$>0.05). In addition, we tested whether the participants interacted differently with the data, depending upon the experiment condition. To test this hypothesis, we computed the total time spent for participants in each condition, including time spent interacting with the data presented in each view in Verifi (e.g., Social Network View, Accounts View). We found no significant differences in the amount of time spent overall or in any specific panel on the interface across the three conditions.

| | Control | Confirm | Disconfirm |
|---|---|---|---|
| Accuracy | 60.4% | 73.8% | 63.1% |
| Fake Prediction | 54.1% | 55.0% | 51.9% |

Table 4: User accuracy and Fake prediction across conditions.

### RQ2: Measuring the Impact of Uncertainty

While we did not find significant differences in users' decisions (e.g., accuracy) between experiment conditions, we expect differences in accuracy and fake prediction given uncertainty in cues for each account. Based on the cues in Figure 3 Left, we categorize accounts into two types: Easy and Difficult. These categories are based on how each account scores on the six cues and are independent from users' responses. In this section, we describe regression analysis to analyze the effect of cues and account on users' decision-making. We then present thematic analysis of users' comments regarding their decisions.

**Regression Analysis**: Our results provide evidence that the prevalent factors in users' decision-making were the cues and the accounts. Table 5 provides the log odds ratios for the independent variables by each regression. We observe three findings. First, in general cues have a significant effect on users' fake prediction and accuracy. For the cues, we recoded the responses to indicate whether the cue was used consistent or not (e.g., depending on the direction of the cue relative to fake or real accounts). We find that the opinionated, fear, and social network cues were the most important in explaining correct decisions when used consistently. Alternatively for explaining Fake decisions, we find that log odds ratios align to the cue direction as mentioned in Figure 3. For example, cues Q2, Q4, and Q6 point to the account being fake and we find the log odds ratios above one, although only Q4 and Q6 are statistically significant.

Second, we find that certain accounts had a significant effect on both users' accuracy and fake prediction. This observation implies that some accounts were more difficult and systematically over or under predicted as fake. For example, @MOMENT has a very low log odds ratio for users' accuracy as users overwhelmingly incorrectly predicted @MOMENT, a real-difficult account, as fake (as indicated by its high log odds ratio for fake prediction).

Last, we find that confidence has no significant relationship in explaining accuracy or fake decisions. While there may be a univariate relationship between confidence and user decisions, this may likely be explained through the account level dummy variables as confidence also varied by

| | Dependent Variable | |
|---|---|---|
| **Independent Variable** | **Accuracy** | **Fake** |
| (Intercept) | 0.18** | 0.21** |
| Social Network Cue (Q1) | 2.03*** | 0.99 |
| Tweet Rate Cue (Q2) | 1.24 | 1.06 |
| Fairness Cue (Q3) | 1.30* | 0.74** |
| Entity Cue (Q4) | 1.43* | 1.77*** |
| Fear Cue (Q5) | 1.53*** | 0.90 |
| Opinionated Cue (Q6) | 2.78*** | 2.74*** |
| @ViralDataInc (Fake Easy) | 9.86*** | 117.96*** |
| @NationalFist (Fake Easy) | 1.90 | 9.7*** |
| @GothamPost (Real Easy) | 0.95 | 0.84 |
| @Williams (Real Difficult) | 0.90 | 2.13* |
| @MOMENT (Real Difficult) | 0.36** | 5.70*** |
| @BYZBrief (Fake Difficult) | 4.91*** | 24.21*** |
| @ThirtyPrevent (Fake Difficult) | 3.70** | 18.89*** |
| User Confidence | 1.14 | 0.86 |
| Confirm Group | 1.97** | 0.91 |
| Disconfirm Group | 1.16 | 0.75 |

*** = 99%, ** = 95%, * = 90% confidence

Table 5: Log odds ratios for each independent variable in two logistic regressions. The Accuracy column is 1 = Correct, 0 = Incorrect Decision. The Fake column is the user's prediction: 1 = Fake, 0 = Real. The @accounts variables use @XYZ as the reference level and the Group variables use the Control Group as the reference level.

accounts. Also, we find the Confirm condition maintains a weakly significant effect on accuracy relative to the Control group (reference level for treatments).

**Thematic Analysis of Comments:** Our regression analysis revealed that cues played an important role in users' decision making on misinformation. When cues point to conflicting directions of an account being real or fake, users are more likely to arrive at inaccurate decisions. In each decision, users had the option to leave comments in regards to their decisions. These comments are extremely valuable in helping us decipher users' rationales. We examined all comments (95 total) and thematically categorized users' strategies. Our analysis focuses on how different usage on all or a subset of the cues affect their decision making. Similar to our quantitative analysis, we evaluate these themes through the lens of cue uncertainty and account difficulties.

Our thematic analysis classified comments into three categories: *Quantitative (32 comments), Qualitative (37), and Qualitative + Quantitative (26). We categorized mentions of social network connection, language feature score, and tweet timeline as quantitative. Any mention related to entities and users' understanding of the text of tweets such as "opinionated language," "news-like text," and "style of text" were considered qualitative. The quantitative and qualitative dimensions extracted from the comments aligned well with the six cues provided to the participants.*

**Easy Accounts**: Easy accounts (column 1, 2, 5, 6 in Figure 3 left) are the ones with most cues pointing to the accounts being either real or fake; thus leading many users to correct decisions. Fifteen comments for the easy accounts mentioned quantitative cues such as language features scores (Figure 4, row 1) and social network connections (Figure 4, row 6) as the basis of their decisions. 12 of these comments led to correct decisions. Seventeen comments focused on the qualitative cues such as opinionated language or entities, e.g., one real account decision based on "factual reporting" and a fake account decision due to seeming "too opinionated" (Figure 4, rows 2 and 5).

**Difficult Accounts**: Difficult accounts (column 3, 4, 7, 8 in Figure 3 left) are the ones with the cues pointing to contradicting directions, resulting in more uncertainties in decision making. Seventeen comments focusing on quantitative cues such as fewer social network connections to other real news accounts for some real-difficult accounts yielded eleven inaccurate decisions (Figure 4, rows 12 and 15). Furthermore, twenty comments focused on qualitative cues such as users' notion of opinionated language, in which seven cases it drove them to wrong decisions (Figure 4, row 11). Finally, fourteen comments focused on both quantitative and qualitative cues with only three of them yielding wrong decisions. In two of these cases, users decided to disregard the account's anger ranking (Figure 4, row 16).

We observe that when users leverage both quantitative and qualitative cues with a thorough analysis of an account, they are more likely to make an accurate decision. Most comments contained a mix of qualitative and quantitative analysis (including language features, social network connections, and opinionated language) helped users to come to the correct decisions (Figure 4, rows 3, 4, 7 and 8).

## Discussion and Future Work

Our goal was to assess the effect of confirmation bias and uncertainties on the investigation of misinformation using visual analytics systems. Although our post-questionnaire vignette, based on prior psychology research (Nickerson 1998) showed that most of our users demonstrated a high level of confirmation bias, our experiment did not find significant differences between the experiment conditions. One explanation would be the hypothesis (all eight accounts are fake) we gave the participants did not resonate with them. If we had asked the participants to form their own hypothesis of the eight accounts being either real or fake by going through an example account, they may have been more invested in the hypothesis and inclined to confirm or disconfirm it. Another explanation involves the use of Verifi, the visual analytics system that empowers users' decision-making by allowing users to interatively analyze multiple aspects of the news accounts. Often, people are instructed to 'slow down' and inspect information more critically (Kahneman 2011) as an antidote to falling for confirmation bias. The Verifi interface could have played a role of somewhat mitigating confirmation bias in our experiments. This will be the subject of our follow-up studies.

We observe that participants' responses to the cues were consistent with the account uncertainties/difficulties. Figure 3-Right shows how users' average cue responses matched

| | Correct? | Type | Group | Comment | category |
|---|---|---|---|---|---|
| 1 | Yes | real | easy | Several language features are consistent with predominantly real accounts | quantitative |
| 2 | | | | News appears more factual reporting rather than opinionated discussion of events, which leads me to believe it is a real news account. | quantitative |
| 3 | | | difficult | This account does not seem to deal much with controversial topics, and although it has a lower loyalty score, it has a high fairness score and high bias, which are normally indicative of real accounts. | quantitative + qualitative |
| 4 | | | | While this account only has one connection and it's to a fake account, I didn't notice anything suspicious in the tweets. The People and Organizations view only showed topics that are normally discussed in the news and nothing overly controversial. | quantitative + qualitative |
| 5 | | fake | easy | A lot of the tweets were not even news but simply them stating their opinions about a variety of issues. | qualitative |
| 6 | | | | Only follows one account, tends to only tweet about one topic (Trump), and it's all negative and uses opinionated language. | quantitative |
| 7 | | | difficult | High fairness but low loyalty. Little amount of tweets (seemed inconsistent). Very high anger. When looking at the network, it was associated with a wide range of different accounts. | qualitative + qualitative |
| 8 | | | | This account is 100% angry, with a low tweet amount. This user also doesn't focus on that many people within their tweets. | quantitative + qualitative |
| 9 | No | real | easy | Compared timeline of tweets as other tweets. The timeline and tweet content about taking Mosul for this account do not match with other "real" news. | quantitative |
| 10 | | | | For this account, language within the tweets tipped me to believing this is a fake news account, or at least an extremely conservative or right-leaning (with high bias) news account. Wordage like "marxist left mainstream media" for instance. | qualitative |
| 11 | | | difficult | Contains a lot of opinionated language in it's tweets. | qualitative |
| 12 | | | | ..Despite the high tweet rate, their bias and subjectivity scores were high, which tends to relate to fake accounts. That added to the fact that it's only linked to another fake account and some verified accounts led me to believe this is a fake. | quantitative |
| 13 | | fake | easy | Though this is very opinionated, it leans towards an overall criticism of America, as opposed to an organization attempting to sway a constituency. | qualitative |
| 14 | | | | Admittedly, personal bias played a role in deciding the "real"ness of this account as the information in the tweets, though not seemingly produced by big media, appears real, though not unbiased. | qualitative |
| 15 | | | difficult | Connected to real accounts and has lower subjectivity. | quantitative |
| 16 | | | | Although there was a high rating of anger, it seems as though none of the tweets expressed any anger or high bias. | quantitative + qualitative |

Figure 4: A sample of users' comments about their decisions. Highlighted text shows users' mention of either a qualitative or quantitative reason. Green denotes reasons/cues pointing to the account being real while red pointing to being fake.

our original understanding of these accounts. Moreover, our regression analysis shows that certain cues significantly affected our users' decisions (Q4-Q6) more than others. Opinionated language which had the strongest effect on users fake prediction stands out as an important lesson learned for future attempts to address misinformation. The fact that we allowed the opinionated cue to be purely based on users' understanding of tweet texts, opens a whole new research question: How can we help users' to more objectively identify/quantify opinionated language?

Furthermore, we find that uncertainty affected our users' prediction accuracy. Our research shows that when a combination of quantitative and qualitative cues are presented clearly and with minimal uncertainty, users are successful in correctly differentiating between fake and real news accounts. In order to be resilient to these uncertainties, it is essential to take effective measures to communicate these uncertainties, motivate users to not be anchored on specific cues, and to holistically focus on a combination of qualitative and quantitative evidence. We plan to conduct a followup experiment with adding uncertainty of the cues to the visual analytic system to test this hypothesis. One limitation of our study was the number of accounts chosen. Due to the time duration of our study (one hour), we decided to ask each participant to make decisions about eight accounts with varying difficulties. In order to test whether our results can be generalized, we plan to conduct a follow-up study that focuses on annotating a larger number of randomized accounts. The current study provides guidance on how we would instruct human coders to categorize all accounts based on the cues into different difficulty levels.

## Conclusion

This paper introduces a visual analytics system, Verifi, along with an experiment to investigate how individuals make decisions on misinformation from Twitter news accounts. We found that the account difficulty as mixed cues indicating real versus fakeness has a significant impact on users' decisions. The Verifi system is the first visual analytics interface designed to empower people in identifying misinformation. Findings from our experiment inform the design of future studies related to decision-making around misinformation aided by visual analytics systems.

## Acknowledgements

The research described in this paper is part of the Analysis in Motion Initiative at Pacific Northwest National Laboratory (PNNL). It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

## References

Adams, W. C. 1986. Whose lives count? tv coverage of natural disasters. *Journal of Communication* 36(2):113–122.

Allan, M. 2017. Information literacy and confirmation bias: You can lead a person to information, but can you make him think?

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–236.

Cho, I.; Wesslen, R.; Karduni, A.; Santhanam, S.; Shaikh, S.; and Dou, W. 2017. The anchoring effect in decision-

making with visual analytics. In *Visual Analytics Science and Technology (VAST), 2017 IEEE Conference on*.

Entman, R. M. 2004. *Projections of power: Framing news, public opinion, and US foreign policy*. University of Chicago Press.

Entman, R. M. 2007. Framing bias: Media in the distribution of power. *Journal of communication* 57(1):163–173.

Evans, J. S. B. 1989. *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96(5):1029.

Haidt, J., and Graham, J. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20(1):98–116.

Hassan, N.; Sultana, A.; Wu, Y.; Zhang, G.; Li, C.; Yang, J.; and Yu, C. 2014. Data in, fact out: automated monitoring of facts by factwatcher. *Proceedings of the VLDB Endowment* 7(13):1557–1560.

Howell, L., et al. 2013. Digital wildfires in a hyperconnected world. *WEF Report* 3:15–94.

John, O. P.; Donahue, E. M.; and Kentle, R. L. 1991. The big five inventory-versions 4a and 54.

Jonas, E.; Schulz-Hardt, S.; Frey, D.; and Thelen, N. 2001. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology* 80(4):557.

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Kott, A.; Alberts, D. S.; and Wang, C. 2015. War of 2050: a battle for information, communications, and computer security. *arXiv preprint arXiv:1512.00360*.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, 342–351. ACM.

Mele, N.; Lazer, D.; Baum, M.; Grinberg, N.; Friedland, L.; Joseph, K.; Hobbs, W.; and Mattsson, C. 2017. Combating fake news: An agenda for research and action.

Metaxas, P. T.; Finn, S.; and Mustafaraj, E. 2015. Using twittertrails. com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, 69–72. ACM.

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2):175.

Rajsic, J.; Wilson, D. E.; and Pratt, J. 2015. Confirmation bias in visual search. *Journal of experimental psychology: human perception and performance* 41(5):1353–1364.

Recasens, M.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2013. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, 1650–1659.

Resnick, P.; Carton, S.; Park, S.; Shen, Y.; and Zeffer, N. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*.

Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 745–750. International World Wide Web Conferences Steering Committee.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*.

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):22–36.

Soll, J. 2017. The long and brutal history of fake news. *POLITICO Magazine*.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, 230–239.

Tsang, A., and Larson, K. 2016. The echo chamber: Strategic voting and homophily in social networks. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 368–375. International Foundation for Autonomous Agents and Multiagent Systems.

Volkova, S.; Bachrach, Y.; Armstrong, M.; and Sharma, V. 2015. Inferring latent user properties from texts published in social media. In *AAAI*, 4296–4297.

Volkova, S.; Shaffer, K.; Jang, J. Y.; and Hodas, N. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, 647–653.

Wagner, M.; Fischer, F.; Luh, R.; Haberson, A.; Rind, A.; Keim, D. A.; Aigner, W.; Borgo, R.; Ganovelli, F.; and Viola, I. 2015. A survey of visualization systems for malware analysis. In *EG Conference on Visualization (EuroVis)-STARs*, 105–125.

Wall, E.; Blaha, L.; Paul, C. L.; Cook, K.; and Endert, A. 2017. Four perspectives on human bias in visual analytics. In *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations*.

Wallach, H. 2018. Computational social science $\neq$ computer science + social data. *Commun. ACM* 61(3):42–44.

Wanner, F.; Stoffel, A.; Jäckle, D.; Kwon, B. C.; Weiler, A.; Keim, D. A.; Isaacs, K. E.; Giménez, A.; Jusufi, I.; Gamblin, T.; et al. 2014. State-of-the-art report of visual analysis for event detection in text data streams. In *Computer Graphics Forum*, volume 33.

Wason, P. C. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology* 12(3):129–140.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 347–354. Association for Computational Linguistics.