I can’t Believe it’s not Better:

Detecting and Quantifying Misinformation and Disinformation

Svitlana Volkova, PhD

Senior Scientist, Data Sciences and Analytics Group
National Security Directorate
Around two-thirds of U.S adults get news from social media 26% get news from at least two social media platforms

Source: News Use Across Social Media Platforms 2018 Pew Research Center – survey conducted July 30-Aug 12, 2018
In 2017, a panel of 50 experts listed the breakdown of trusted information sources as one of the grand challenges of the 21st century.


Massive digital disinformation is one of the main risks of the modern society.

Misinformation and Disinformation Trends According to Google

Source: https://trends.google.com/trends/explore?date=all&q=misinformation,disinformation
Trust to Social Media News

2016

64%
of adults believe fake news stories cause a
great deal of confusion

23% said they had shared fabricated stories
themselves – by mistake or intentionally

2018

31% of social media news consumers list
inaccuracy as their top concern

Largely inaccurate: 57%
Largely accurate: 42%

News Use Across Social Media Platforms 2018, Pew Research Center – survey conducted July 30-Aug 12, 2018

I can't Believe it's not Better: Detecting and Quantifying Misinformation and Disinformation

March 1, 2019
In 2017, 114 Active Fact-checking Services: 19% Increase from 2016

Fact-checking after the spread is ineffective
Surviving in the Post-truth Era

**Post-truth** is relating to circumstances in which **objective facts** are less influential in shaping public opinion than **appeals to emotion and personal beliefs**

This talk will focus on multi-dimensional computational approaches to disinformation and misinformation

I. Identifying deceptive news types
   ▪ Multilingual
   ▪ Multimodal

II. Measuring the immediate spread of deceptive news
   ▪ Who shares, how evenly and how quickly

III. Quantifying reactions to deceptive news
   ▪ Bots versus humans
This talk will focus on multi-dimensional computational approaches to disinformation and misinformation

I. Identifying deceptive news types
   - Multilingual
   - Multimodal

II. Measuring the immediate spread of deceptive news
   - Who shares, how evenly and how quickly

III. Quantifying reactions to deceptive news
   - Bots versus humans
Misinformation and Disinformation

- **Trusted News**: No desire to deceive, trustworthy information.
- **Propaganda**: To deceive, false information.
- **Clickbait**: To deceive, false information.
- **Conspiracy Theories**: To deceive, false information.
- **Hoax**: To deceive, false information.
- **Satire**: No desire to deceive, false information.
- **Disinformation**: To deceive, false information.

I can't Believe it's not Better: Detecting and Quantifying Misinformation and Disinformation

March 1, 2019
Disinformation refers to the deliberate planting of false reports

- Disinformation in a report on intelligence activities leading up to World War II in 1939

Misinformation equates in meaning but does not carry the same devious connotation

Burned to death because of a rumour on WhatsApp

By Marcos Martínez
BBC Monitoring

© 12 November 2018


Fighting The Endless Spread Of Ebola Misinformation On Social Media

This is how an epidemic unfolds online: with rumors, scientific fact, and prayers all mixed together.

Misinformation Types

- **Propaganda**: Art of persuading, influencing and manipulating public opinions, attitudes, and actions.
- **Conspiracy**: Effort to explain an event or practice by the coordinated actions of powerful people.
- **Hoax**: Scams or deliberately false or misleading stories made up for financial or political gain.
- **Clickbait**: Flashy, engaging content for attracting web traffic.
- **Satire**: Misleading statements with the primary purpose to entertain.
- **No Desire to Deceive**: I can’t Believe it’s not Better: Detecting and Quantifying Misinformation and Disinformation.
Predicting Deceptive News Types

Model Inputs

- Twitter data in English from 2016
  - 134 deceptive news accounts\(^1\)
  - 166 trusted accounts
- Twitter samples:
  - 130K posts
  - 2M posts
  - 7M posts in English and other languages (Russian, German, French) from news accounts

Predictive Tasks

- Binary: deceptive, trusted
- 4-way: propaganda, clickbait, hoax, satire
- 5-way: the above + disinformation

\(^1\)Preexisting lists of deceptive news sources constructed by fact-checkers, journalists, and academics: fact-checking, hoaxy, news watch and disinfo


I can't Believe it's not Better: Detecting and Quantifying Misinformation and Disinformation
Linguistic Markers of Deception: Lexical Resources

I. Biased Language
- Assertive, factive, hedging, implicative, and report verbs, comparatives, superlatives, action, manner, and modal adverbs[2]
- As the Communist Party USA website claims...

II. Subjective Language
- Used to dramatize or sensationalize a news story – strongly and weekly subjective terms[3]
  - He has one of the most brilliant minds in basketball...

III. Moral Foundations
- People differ in the way they endorse moral foundation values[1]
  - Care/Harm: kindness and gentleness
  - Fairness/Cheating: justice, rights, and autonomy
  - Loyalty/Betrayal: patriotism and self-sacrifice for the group

Linguistic Markers for Deception: Twitter

How frequently do lexical markers appear in deceptive posts compared to the trusted?

- Weak. Subj. 6.0
- Care 3.2
- Strong. Subj. 1.8
- Hedging 1.5
- Implicative 1.3
- Harm 0.7
- Report 4.5
- Betrayal 1.4
- Factive 0.6
- Loyalty 0.4
- Authority 0.5

*All findings are statistically significant, p < 0.001

Linguistic Markers for Deception: News Articles

How frequently do lexical markers appear in deceptive posts compared to the trusted?

<table>
<thead>
<tr>
<th>Marker Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swear (LIWC)</td>
<td>7.00</td>
</tr>
<tr>
<td>Modal Adverb</td>
<td>2.63</td>
</tr>
<tr>
<td>Action Adverb</td>
<td>2.18</td>
</tr>
<tr>
<td>1st pers singular</td>
<td>(I) 2.06</td>
</tr>
<tr>
<td>Manner Adverb</td>
<td>1.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marker Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd pers (You)</td>
<td>6.73</td>
</tr>
<tr>
<td>Sexual (LIWC)</td>
<td>1.80</td>
</tr>
<tr>
<td>See (LIWC)</td>
<td>1.52</td>
</tr>
<tr>
<td>Weak subjective</td>
<td>1.13</td>
</tr>
<tr>
<td>Strong subjective</td>
<td>1.51</td>
</tr>
<tr>
<td>Negation (LIWC)</td>
<td>1.51</td>
</tr>
<tr>
<td>Hedge</td>
<td>1.19</td>
</tr>
<tr>
<td>Superlatives</td>
<td>1.17</td>
</tr>
<tr>
<td>Assertive</td>
<td>0.84</td>
</tr>
<tr>
<td>Money (LIWC)</td>
<td>0.57</td>
</tr>
<tr>
<td>Comparitives</td>
<td>0.86</td>
</tr>
<tr>
<td>Hear (LIWC)</td>
<td>0.50</td>
</tr>
<tr>
<td>Number (LIWC)</td>
<td>0.43</td>
</tr>
</tbody>
</table>

*All findings are statistically significant, p < 0.001
Predicting Deceptive vs. Trusted News

- Text + Network + Ling. Markers yields the best performance

Small English Dataset (balanced), 130K posts

Predicting Deceptive News Types: Multiclass

- Text + Social Network Interactions yield the best results
- Ling. Markers do not distinguish deceptive classes

Small English Dataset, 130K posts


I can't Believe it's not Better: Detecting and Quantifying Misinformation and Disinformation
Predicting Deceptive News Types: Additional Data

- More data → better performance
- Text < Text + Network
Predicting Deceptive News Types: Additional Classes

- Text + Network $\gg$ Text + DeepWalk$^1$

Predicting Deceptive News Types: Additional Languages

Large Multilingual Dataset, 7M posts from 2016
Predicting Deceptive News Types: Additional Languages

• Unlike for English, characters in combination with DeepWalk representations is the best

• Social interactions of EN-speaking and RU-speaking users with deceptive news sources\(^1\) is different

Multimodal Deception Prediction

Prediction Tasks

- 4-way: conspiracy, clickbait, hoax, satire (2.5K posts)
- 2-way: deceptive (propaganda, disinfo) vs. trusted, 37K posts
- 3-way: propaganda, disinfo, trusted, 56K posts


I can't Believe it's not Better: Detecting and Quantifying Misinformation and Disinformation

March 1, 2019
Multimodal Deceptive News Classification

• Text + Image + Ling. Markers yield the best performance for fine-grained prediction (4-way)
F1 Scores are Great! But … How about Explaining Model Predictions and Contrasting Models?
Representative Deceptive Images Per Class

**Clickbait:** head shots of politicians or celebrities

**Disinformation:** politicians and celebrities overlaid with text

**Propaganda:** natural scenes

**Satire:** individuals but not politicians

**Graphs and charts are highly indicative of conspiracy**

**Hoaxes:** pictures of newspapers or magazine articles

---

I can't Believe it's not Better: Detecting and Quantifying Misinformation and Disinformation

March 1, 2019
This talk will focus on multi-dimensional computational approaches to disinformation and misinformation

I. Identifying deceptive news types
   - Multilingual
   - Cross-domain
   - Multimodal

II. Measuring the immediate spread of deceptive news
   - Who shares, how evenly and how quickly

III. Quantifying reactions to deceptive news
   - Bots versus humans
Sharing Inequality: How Evenly Users Share Deceptive vs. Verified News

Does each user share equally? OR Does a small group do the most re-regrees?

10M retweets or mentions from deceptive and verified news sources
1.7M unique users

I can't Believe it's not Better: Detecting and Quantifying Misinformation and Disinformation
Measures of Participation Inequality

Disparity in Re-sharing Behavior across Deceptive News Types

How unequal is participation?

- Disinformation sources are most highly retweeted from a small group of users

[Graph showing the disparity in re-sharing behavior across different news types]

News Propagation and Influence from Deceptive Sources, M. Glenski, T. Weninger and S. Volkova. TCSS journal.
Inferring User Demographics

- 145K users with ≥ 5 interactions with deceptive sources (English only)
- Classify users as person or organization using the Humanizr
- 66K unique person-users with ≥ 200 public tweets

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Inferred Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>M : Male</td>
</tr>
<tr>
<td><strong>AUC 0.89</strong></td>
<td>F : Female</td>
</tr>
<tr>
<td>Age</td>
<td>Y : 24 and younger</td>
</tr>
<tr>
<td><strong>AUC 0.72</strong></td>
<td>O : 25 and older</td>
</tr>
<tr>
<td>Income</td>
<td>B : Below $35k</td>
</tr>
<tr>
<td><strong>AUC 0.72</strong></td>
<td>A : At least $35k</td>
</tr>
<tr>
<td>Education</td>
<td>H : High School</td>
</tr>
<tr>
<td><strong>AUC 0.76</strong></td>
<td>D : At least Bachelors</td>
</tr>
</tbody>
</table>

*manually annotated data from Volkova et al., 2016

**News Propagation and Influence from Deceptive Sources**, M. Glenski, T. Weninger and S. Volkova. TCSS journal.
Who is more likely to share at least once?
Who is more likely to share more of each type?

Gender
- Trusted
- Clickbait
- Conspiracy
- Propaganda
- Disinformation

Age
- Trusted
- Clickbait
- Conspiracy
- Propaganda
- Disinformation

Income
- Trusted
- Clickbait
- Conspiracy
- Propaganda
- Disinformation

Education
- Trusted
- Clickbait
- Conspiracy
- Propaganda
- Disinformation

p < 0.01

Participation Inequality with Palma Ratios: Inferred Gender and Age

**Women** retweet conspiracy theory more evenly than **men**

The most active 10% of **younger** users who retweet disinformation sources share 41.34 times what the least active 40% do.

*News Propagation and Influence from Deceptive Sources*, M. Glenski, T. Weninger and S. Volkova. TCSS journal.
Speed of the Immediate Deception Spread

- Science paper by Vosoughi et al., 2018 reports "False news spread more rapidly than trusted" [1]
- Users spread faster information from conspiracy, disinformation or verified sources than from clickbait and propaganda
- Older users propagate news from trusted sources more quickly than younger, but they share from suspicious sources after longer delays

This talk will focus on multi-dimensional computational approaches to disinformation and misinformation

I. Identifying deceptive news types
   ▪ Multilingual
   ▪ Multimodal

II. Measuring the immediate spread of deceptive news
    ▪ Who shares, how evenly and how quickly

III. Quantifying reactions to deceptive news
    ▪ Bots versus humans

Predicting Bot Scores and Reaction Types

I. 431K English tweets, 184K users, 255 news sources
   - Less that 10% bots, 70% humans, and 20% unknown

II. Reaction model\(^1\) learned from 83K manually labeled Reddit comments

---

Bot annotations are coming from Characterizing online discussion using coarse discourse sequences. A. Zhang, B. Culbertson, and P. Paritosh. NAACL 2017.

Reaction Variety: Distributions within sub-populations

Conspiracy:
% Question > % Answer  (Human & Bot)

Propaganda:
% Questions > % Answers (Human)
% Answers > % Questions (Bot)

- Bots↓:
  - Answer/Question ➥ Conspiracy Elaboration ➥ Propaganda

- Bots↑:
  - Answer ➥ Propaganda/Clickbait Elaboration ➥ Disinformation Question ➥ Clickbait/Disinformation

---

Reaction Speed: Bots React Slower than Humans except to Propaganda and Clickbaits

• Bots have a higher level of disparity in reaction volumes than humans

Lessons Learned: Identifying Deceptive News Types

I. Identifying deceptive news types
   - Multilingual
   - Multimodal

- Combining Text + Network + Ling. Markers is the best
- Linguistic markers are more helpful for separating deceptive vs. trusted language rather than across deception types
- Multilingual predictions are harder
- Incorporating image representations in addition to other signals is beneficial

What’s next? Explaining model prediction and analyzing errors
Lessons Learned:
Measuring the Spread of Deceptive News

• **Who shares:** User predicted to be younger, less educated, female, higher incomer are likely to share disinformation (users with opposite attributes share misinformation more)

• **How evenly:** Disinformation is most highly re-shared from a small group of users

• **How quickly:** Users spread faster news from conspiracy, disinformation or verified sources than from clickbait and propaganda

II. Measuring the immediate spread of deceptive news
   - Who shares, how evenly and how quickly

• **What’s Next?** DARPA SocialSim program focuses on:
  - Measuring cross-platform information spread and evolution
  - Scope and scale of information spread (including mis/disinfo)
  - Spread phenomena: information cascades, recurrence, gatekeepers
Lessons Learned: Quantifying Reactions to Deceptive News

I. Identifying deceptive news types

§ Multiple domains
§ Multilingual
§ Multimodal

II. Measuring the immediate spread of deceptive news

§ Who shares, how much, how evenly and how quickly

III. Quantifying reactions to deceptive news

§ Bots versus humans

• Reaction variety: Humans question propaganda more often than they provide answers but bots question less

• Reaction speed: Bots react slower than humans except when answering to Clickbait and Propaganda

• Reaction inequality: Bots have a higher level of disparity in reaction volumes than humans

• What’s next? Reactions across languages, locations and social platforms
Research Highlights on Mis/Disinformation

Confirmation bias and user studies

Deception spread

Reactions to deceptive news
- **Identifying and Understanding User Reactions to Deceptive and Trusted Social News Sources.** M. Glenski, T. Weninger and S. Volkova. ACL 2018

Identifying types of deceptive news
- **Misleading or Falsification? Inferring Deceptive Strategies and Types in Online News and Social Media.** S. Volkova and J. Jang. WWW Track on Journalism, Misinformation and Fact Checking 2018.
The spreading of misinformation online

Influence of fake news in Twitter US presidential election

The dynamics and influence of fake news on Twitter during the 2016 US presidential election remains to be clarified. Here, we use a dataset of 171 million tweets in preceding the election day to identify 30 million tweets, from 2.2 million users, that link to news outlets. Based on a classification of news outlets curated by the Washington Post, we find that 25% of these tweets spread either fake or extremely biased news. We characterize the network of information flow to find the most influential fake news outlets and traditional news and use causal modeling to uncover how fake news spread.

Finding facts about fake news

There was a proliferation of fake news during the 2016 election cycle. Grinberg et al. analyzed Twitter data by matching Twitter accounts to specific voters to determine who was exposed to fake news, who spread fake news, and how fake news interacted with factual news (see the Perspective by Christo). We find that heavily concentrated in specific locations and are responsible for the spread of fake news, particularly polarizing and conservative news. Science, this is an issue that needs to be addressed urgently.
Confirmation Bias

https://verifi.herokuapp.com


Vulnerable to misinformation? Verifi IUI 2019.

People tend to seek information that aligns with their views
In the battle against misinformation, it is better to prevent that cure.


Misinformation is not like a plumbing problem you fix. It is a social condition, like crime, that you must constantly monitor and adjust.

Tom Rosenstiel, journalist, researcher, media critic

How can we create a news ecosystem and culture that value and promotes truth?

Lazer at al, Science 2018

Collaborators

Dustin Arendt, PhD
Jin Yea Jang
Post-Masters
Kyle Shaffer
Data Scientist
Josh Mendoza,
Data Engineer
Maria Glenski
PhD Student, UND
Tim Weninger, PhD
UND
Ryan Wesslen
PhD Student, UNCC
Wenwen Dou, PhD
UNCC
Mark Greaves, PhD
Emily Grace, PhD
I can’t Believe it’s not Better: Detecting and Quantifying Misinformation and Disinformation
Ellyn Ayton
MS Student, WWU
Brian Hutchinson, PhD
WWU
Hannah Rashkin
PhD Student, UW
Yejin Choi, PhD
UW
Hannah Rashkin
PhD Student, UW
Tim Weninger, PhD
UND
Yejin Choi, PhD
UW
Wenwen Dou, PhD
UNCC
Mark Greaves, PhD
Emily Grace, PhD
I can’t Believe it’s not Better: Detecting and Quantifying Misinformation and Disinformation

March 1, 2019
Thank you

svitlana.volkova@pnnl.gov
https://www.cs.jhu.edu/~svitlana/

Looking for interns and postdocs. If interested, please email me.