Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background

Kevin Roitero, **Michael Soprano**, Shaoyang Fan,
Damiano Spina, Stefano Mizzaro and Gianluca Demartini

ACM SIGIR 2020

Context

- The information we are exposed to influences our decision making processes
- Understanding which information should be trusted/untrusted is fundamental for democracy processes to function
- Information credibility assessment (***fact checking***) is a task that has gained popularity due to the spread of misinformation online (expert fact checkers)
- The volume of misleading and false information increases [Nguyen and Kyumin 2018]
- Expert fact checkers cannot handle such volume of misinformation

[Nguyen and Kyumin 2018] Nguyen Vo, & Kyumin Lee. (2018). The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News.

Aims

- Study **limitations of non-expert fact checkers identifying misinformation online**
- Very large crowdsourcing experiment
    - Each crowd worker is asked to fact check statements given by politicians

Research Questions

- **RQ1**: Suitability of different assessment scale to gather truthfulness labels
- **RQ2**: Relationship between crowd and expert truthfulness labels
- **RQ3**: Sources of information that crowd workers use to identify online misinformation
- **RQ4**: Effect and role of assessors' background in identifying online misinformation

Dataset

Two sets of statements made by politicians

- **Politifact [Yang 2017]**:
  - 12,800 statements, US politicians
  - Statements annotated by expert fact checkers
  - Truthfulness label on a six-level scale
- **ABC[1]**:
  - 407 statements from 2013 to 2015, Australian politicians
  - Statements annotated by expert fact checkers
  - Fine grained labels aggregated in a three-level scale

[Yang 2017] William Yang Wang. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection.
[1]https://apo.org.au/collection/302996/rmit-abc-fact-check

Assessment Scales and Collections

We consider 3 different truthfulness scales and 5 generated collections

- **2 sets of ground truthfulness labels** by experts
  - Politifact: six-level scale
    - pants-on-fire, false, barely-true, half-true, mostly-true and true
  - ABC: three-level scale
    - negative, in-between, positive
- **3 sets of truthfulness labels** created by means of our task
  - S_6: same as Politifact
  - S_3: same as ABC
  - S_{100}: a 101 level scale in the [0, 100] range

Crowdsourcing Task Design

- 120 (Politifact) + 60 (ABC) = **180 statements** judged by 10 distinct crowd workers
- 20 statements for each ground truth category
- This setting is repeated over **3 different assessment scales** (S_3, S_6 or S_{100})
- Each worker judges 6 (Politifact) + 3 (ABC) + 2 (gold questions) = **11 statements over 1 assessment scale**
- 600 hits over 3 batches (one for each scale)
- A total of 1,800*3=5,400 judgments collected
    - 6,600 considering also gold questions

HITs

- Questionnaire to collect **worker background**
- Three Cognitive Reflection Tests (CTR) to assess **worker cognitive abilities**
  - [Frederick 2005]
- Judgment of **11 statements**
  - 3 Republican, 3 Democratic
  - 1 Labor, 1 Liberal
  - 2 Gold Questions
- The worker must provide:
  - a **truthfulness level** on a given scale (S_3, S_6 or S_{100})
  - **URL** which serves both as **justification** for the judgments and as **source of evidence**
- **Quality checks** to ensure quality of collected data

	Statement	Speaker, Year
PolitiFact Label: mostly-true	"Florida ranks first in the nation for access to free prekindergarten."	Rick Scott, 2014
ABC Label: in-between	"Scrapping the carbon tax means every household will be $550 a year better off."	Tony Abbott, 2014

[Frederick 2005] Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives*, 19 (4): 25-42.

Worker Background

About 6000 US resident crowd workers participated to our study. Across all experiments:

- ~46% of workers are between 26 and 35 years old
- ~61% of workers have a four years college degree at least
- ~68% of workers earn less than $75,000 a year
- ~47% of workers thinks their view are more democratic
- ~58% of workers consider Democratic and Liberal party as their voting preference
- ~52% of workers are again the construction of a wall on the southern border
- ~80% of workers thinks that the government should strengthen environmental regulation to prevent climate change

We have an **overall balanced set of crowd workers**

Worker Behavior

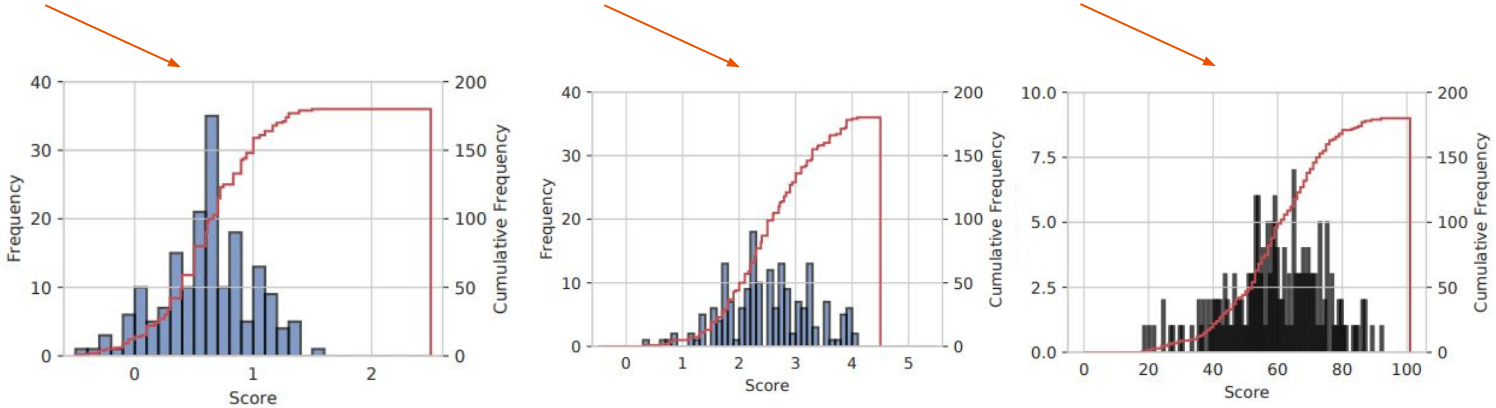
- **Abandonments number are in line with previous studies** [Han 2019]
- Higher (lower) failure (completion) rate for S_{100}
- Workers which go back to previously seen statement are less than 5% over all scales

	Completion	Abandonment	Failure
S_3	35	53	12
S_6	33	52	14
S_{100}	25	53	22

[Han 2019] L. Han et al., "The Impact of Task Abandonment in Crowdsourcing," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2019.2948168.

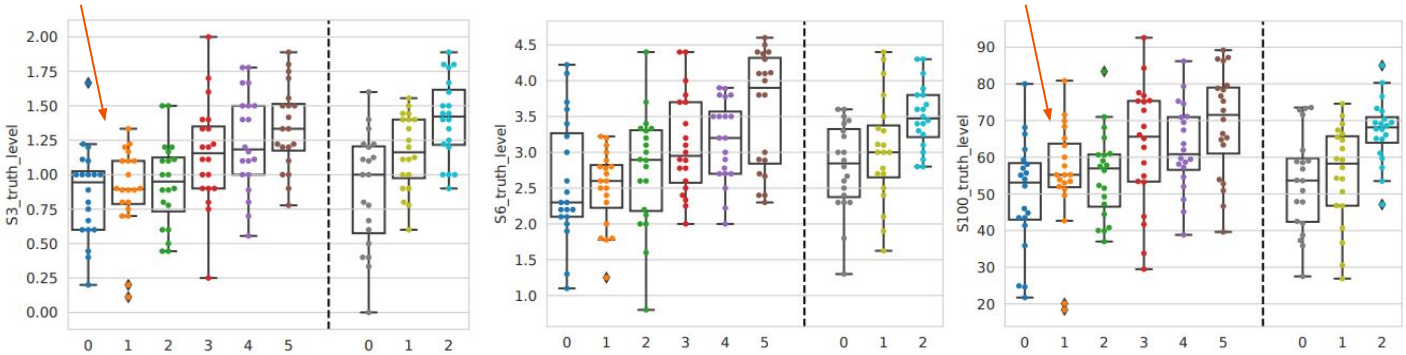
Crowdsourced Scores Distribution

- The 10 scores obtained for each statement are aggregated
- Distribution skewed towards lower values for S_3
- Distribution skewed towards higher values for S_6 and S_{100}
- **Different scales are used differently by crowd workers**



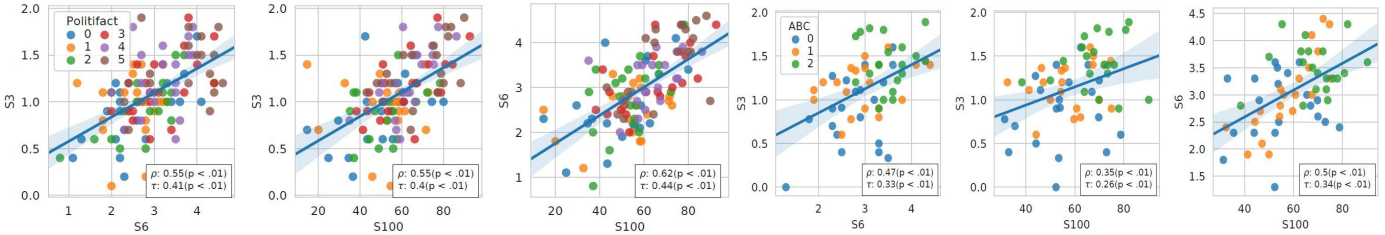
External Agreement

- Agreement between aggregated crowd judgments and ground truth
- Similar behaviour over each scale, both on Politifact and ABC statements
- Politifact: difficult for workers to distinguish between pants-on-fire and false labels
- Higher agreement with the ground truth for higher truthfulness values
    - **Workers recognize true statements more easily than false ones**



Internal Agreement

- Agreement measured among the workers using [Krippendorff 2011] α coefficient
- Rather low agreement among the workers
- We perform all the possible transformations of judgments from one scale to another [Han et al. 2019]
  - **The same statements on different scales tend to be judged differently**
- Overall, across all collections there is a **low level of internal agreement among workers**
  - both within the same scale and across different scales



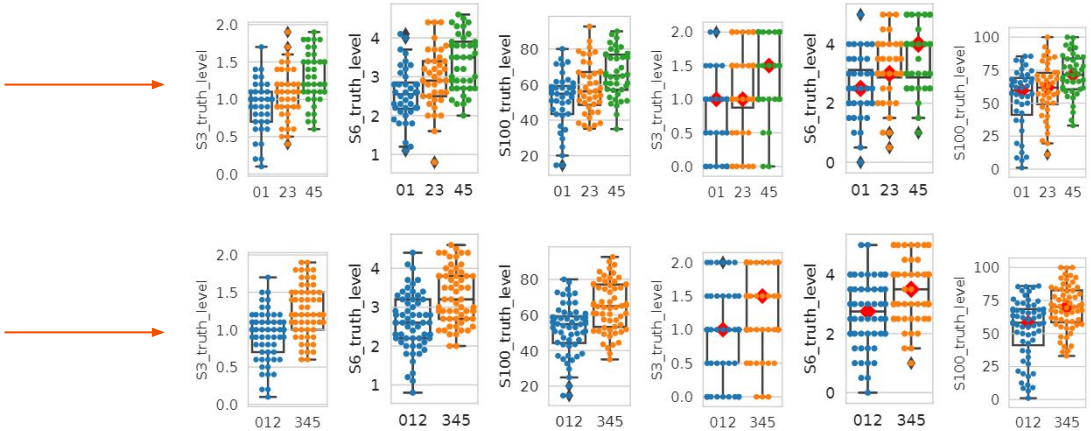
[Han et al. 2019] Han, L., Roitero, K., Maddalena, E., Mizzaro, S., & Demartini, G. (2019). On Transforming Relevance Scales. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 39–48). Association for Computing Machinery.
[Krippendorff 2011] Klaus Krippendorff. Computing Krippendorff's Alpha-Reliability. (2011).

Alternative Aggregation Functions

- We study the effect of using aggregation functions alternative to the arithmetic mean
  - Median
  - Majority vote
- However, **the arithmetic mean is the most effective aggregation function**

Merging Assessment Levels

- We group adjacent categories to check if looking at data on a more coarse-grained ground truth improves the results
  - six Politifact categories in either three (01, 23, 45) or two (012, 345) new categories
  - **the crowd is able to single out true from false statements with good accuracy**



Sources of Evidence

- The most used sources are “Wikipedia” and “Youtube”
- There are also popular news websites and a fact checking one (“FactCheck”)
- **Workers tend to identify trustworthy information sources to support their judgments**
- The majority of workers tend to click on the first result shown by search engine
- Nevertheless, there are also **workers which put some effort to find a reliable source**

	Wikipedia	Youtube	The Guardian	Factcheck	Smh	Cleveland	Washington Post	News	Blogspot	On the Issues
S_3	17	13	11	8	6	6	6	5	5	4
S_6	19	13	12	8	3	6	6	4	6	0
S_{100}	23	12	13	9	6	5	6	5	5	0

	1	2	3	4	5	6	7	8	9	10	11
S_3	17	12	13	14	12	9	7	6	4	3	1
S_6	13	13	16	12	11	9	8	7	5	4	1
S_{100}	15	15	15	12	8	12	7	6	5	2	1
avg	15	13	15	11	10	10	7	6	5	3	1

Effect of Worker Background and Bias

- Relationships between workers' background and their CT performance
- **Workers with strong analytical abilities can better recognize true statements from false**; such ability increases with age
- Workers who have liberal views can better differentiate between false and true statements
- Workers who do not want a wall on the southern US border perform better in distinguishing between true and false statements

Conclusions / Take Home Messages

- **RQ1**: the grouping of adjacent categories reveal that crowdsourced truthfulness judgments are useful to single out true from false statements
- **RQ2**: high external agreement towards higher truthfulness values; low internal agreement between workers
- **RQ3**: workers put effort in finding a reliable source to justify their judgments and tend to choose a source found within the first page of search results
- **RQ4**: assessors' background has an effect on objectively identify online misinformation

Thank You!

Contacts:

- Università degli Studi di Udine, Udine, Italy
  - Kevin Roitero - roitero.kevin@spes.uniud.it
  - **Michael Soprano - soprano.michael@spes.uniud.it**
  - Stefano Mizzaro - mizzaro@uniud.it
- University of Queensland, Brisbane, Australia
  - Shaoyang Fan - fsysean@gmail.com
  - Gianluca Demartini - g.demartini@uq.edu.au
- RMIT University, Melbourne, Australia
  - Damiano Spina - damiano.spina@rmit.edu.au

Data available at https://github.com/KevinRoitero/crowdsourcingTruthfulness