SEMIFORM@ISWC 2020 November 2

Towards Crowdsourcing Tasks for Accurate Misinformation Detection

Ronald Denaux, Flavio Merenda, and
Jose Manuel Gómez-Pérez

Work supported by the European Commission under grant 770302 – Co-Inform – as part of the Horizon 2020 research and innovation programme.

Crowdsourcing

in Misinformation Detection

EXPERT SYSTEM

Can it be done?

Crowdsourced Factchecking

There is a role for crowdsourcing in factchecking but (so far) it's not factchecking

Can crowdsourcing scale fact-checking up, up, up? Probably not, and here's why

NiemanLab

"We foolishly thought that harnessing the crowd was going to require fewer human resources, when in fact it required, at least at the micro level, more."

By **MEVAN BABAKAR** June 6, 2018, 9:42 a.m.

WikiTribune

WikiTribune was a news wiki where volunteers wrote and curated articles about widely publicised news by proof-reading, fact-checking, suggesting possible changes, and adding sources from other, usually long established outlets. Wikipedia

Date launched: April 25, 2017

Type of site: Online newspaper

Headquarters: London

Available in: Spanish Language

Owner: Jimmy Wales

The fact-checking process is highly complex and not amenable to crowdsourcing.

Tradeoff between coverage, complexity and speed.

- Spotting: biased upvoting, long tail ignored
- Finding primary sources: crowd tends to find secondary sources
- Synthesize conclusions & writeup: demotivates participants when high quality required

Political Fact Checking

JOIN

r/politicalfactchecking

PINNED BY MODERATORS

211

Posted by u/CaspianX2 6 months ago

Mod Post/Meta A message to all users of r/PoliticalFactChecking – We are officially closing this subreddit

21 Comments Give Award Share ...

20

Posted by u/NoNoNoThrowaway3002 6 months ago

Most abortions performed at or after 21 weeks are not due to medical reasons.

The data:

About Community

A crowd-sourced effort to monitor and critique the factual accuracy of US political speech. This includes major politicians, political voices, traditional and new media articles, and even email and social network memes. Our goal is to be very specific in checking for well sourced verifiable facts, not evaluating political arguments. And we will strive for as objective and neutral point of view as possible.

12.4k Members

12 Online

Truthsquad™

EXPERT SYSTEM

Can we find a sweetspot?

Misinformation Detection:

- does *this* document contain inaccurate claims?
- If so, which claims and **what evidence is there against them**?

0. If we automate the process...

Extracting sentences

Spotting claims

Match them to primary sources or evidence

Even if not perfectly...

1. Can we derive simple crowdsourcing tasks?

Long primary sources or evidence leading to them are difficult to find

Better to work at smallest level possible: sentence/claim?

This may also limit bias as context is removed

2. Can we improve the system?

Use the feedback to detect errors of initial system

Use feedback to obtain improved system

Go to step 0.

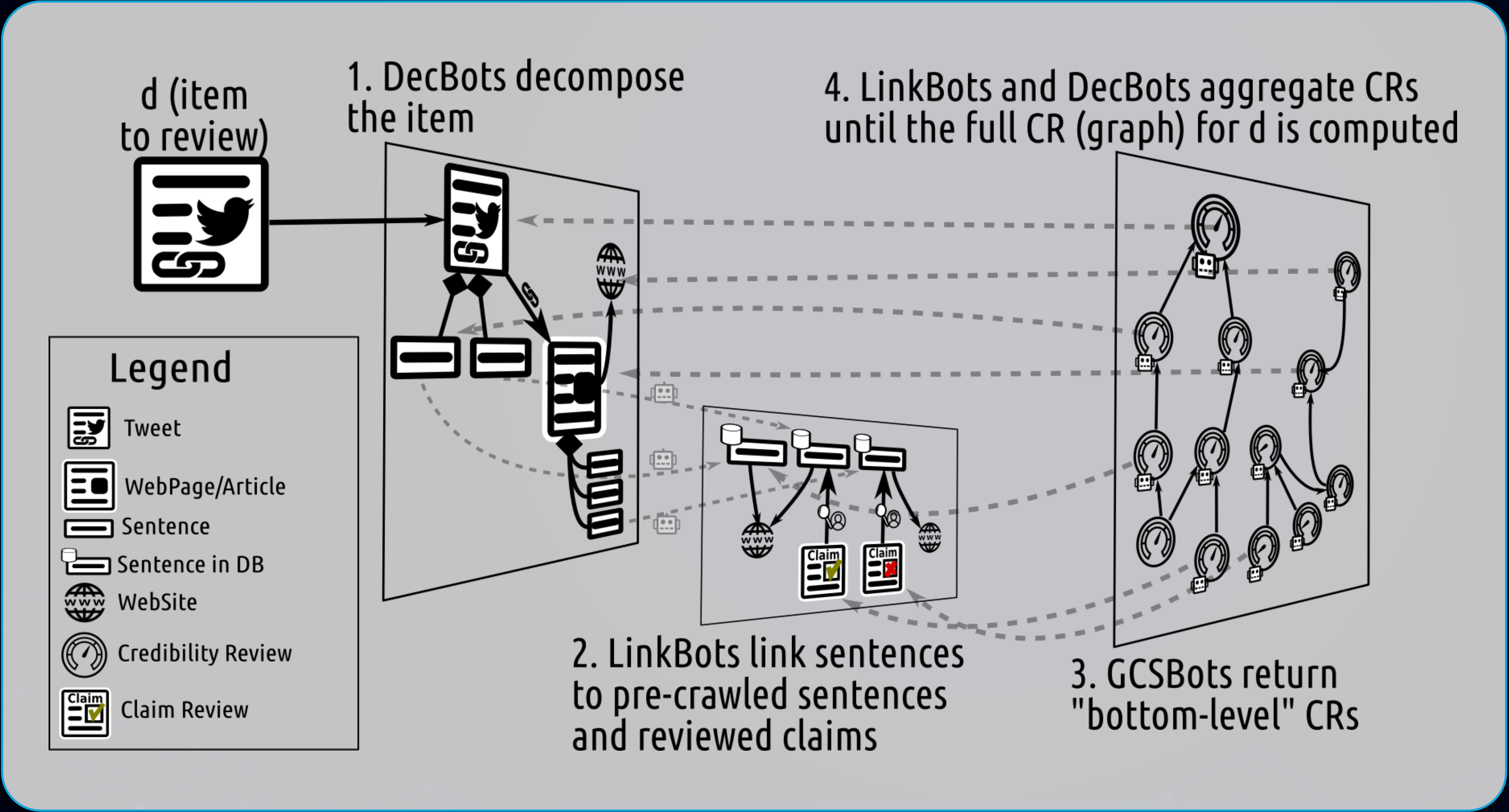
acred & co-inform

Credibility Reviews

for automated misinformation detection

EXPERT SYSTEM

acred



Result: Review Graph

Main Review with:

- credibility rating & confidence
- links to sub reviews and eventually evidence

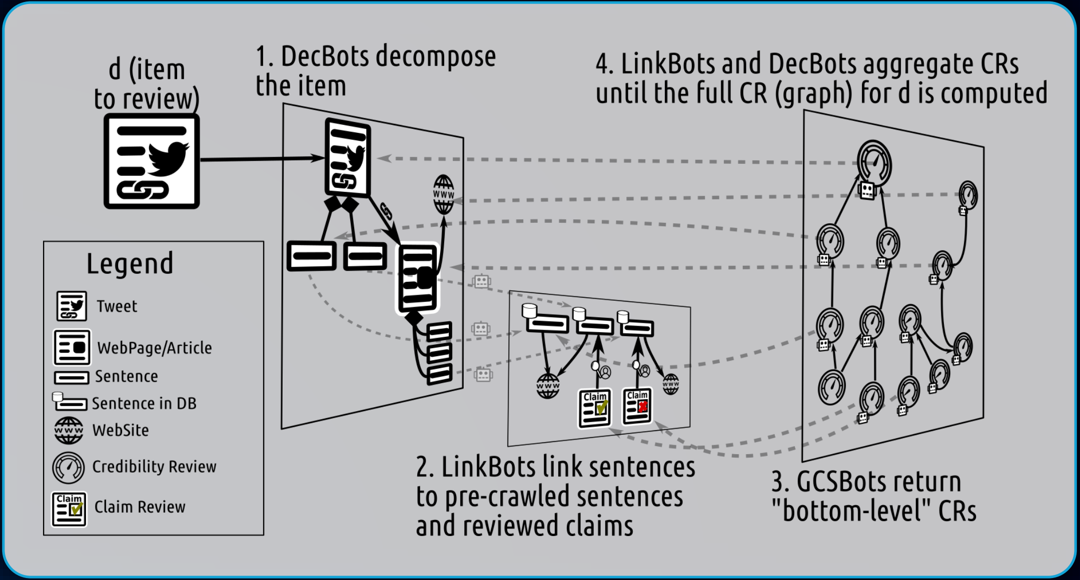
Can be rendered as a label (e.g. "not credible")

See main conference presentation for conceptual and data model

Linked Credibility Reviews

acred

Steps use Deep Learning models, implemented as finetuned RoBERTa instances



Checkworthiness

Is a Sentence a verifiable claim?

Finedtuned on 7.5K samples from CBD (8.7K), Clef'20 Task1 (637) and claims for which a ClaimReview exists (4.6K)

0.85 weighted F1 on Clef'19 test (7K)
0.95 weighted F1 on CB2020 (100)

Semantic Sentence Similarity

How similar are 2 sentences?

Finetuned on STS-B train(5.7K)

0.83 pearson correlation on STS-B dev (1.5K)

Stance Detection

Confirm similarity and provide polarity

Finetuned on FNC-1 train (50K)

92% accuracy on FNC-1 test (25K)

If you have the right data, these models perform really well

EXPERT SYSTEM

acred

Evaluation

FakeNewsNet (Politifact)

420 fake + 528 real webpages/articles
Classes: fake, real

System	Accuracy	Precision	Recall	F1
acred	0.586	0.499	**0.823**	0.622
$acred^{+}$	**0.716**	0.674	0.601	**0.713**
CNN	0.629	**0.807**	0.456	0.583
SAF/S	0.654	0.600	0.789	0.681

72% accuracy

coinform250

250 tweets
Classes: credible, mostly credible, uncertain, not credible, not verifiable

credible	2	0	3	5	0	4
≈cred	3	0	35	8	0	4
uncertain	6	0	41	34	7	22
¬verifiable	2	0	4	4	0	5
≈¬cred	0	0	0	0	0	0
¬credible	7	0	11	21	0	23

27% accuracy

coinform4550 “train”

400 tweets
Balanced co-inform classes
“Silver” labels (automatically mapped from ClaimReviews using MisinfoMe)

	credible	≈cred	uncertain	¬verifiable	¬credible	check_me
credible	18	2	9	37	14	0
≈cred	2	22	3	5	1	0
uncertain	4	1	27	35	13	0
¬verifiable	10	2	20	30	18	0
¬credible	6	0	5	47	22	0
check_me	0	0	0	0	0	0

33% accuracy

EXPERT SYSTEM

acred

Error analysis

Feasible! but time-consuming when finding cause

Possible causes

- Dataset issue: e.g. insufficient content, mislabeled
- Selected non-claim as least credible sentence
- Incorrect semantic similarity + stance leads to:
  - incorrect linking to evidence, or
  - over/underestimating confidence in link
- Incorrect confidence affects aggregation
- Type of evidence: Website Review vs ClaimReview

Example from fakeNewsNet real as “not credible” (sample of 26)

But expecting 92% accuracy!

77% (20) cases involving stance

- 50% (13) stance is overestimated (agree/disagree instead of discuss or unrelated)
- 27% (7) correct stance (unrelated), but not reduced confidence enough

In general

BERT based models great at coarse level, but struggle with specific domains and entities. Can we fix this with additional domain specific samples?

We need a “gold” label to know that we made a mistake and to assess its severity and possible causes.

EXPERT SYSTEM

Crowdacred

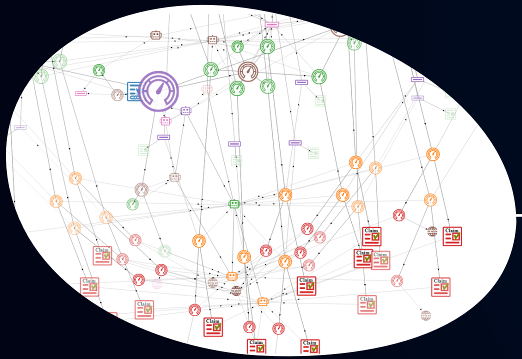
EXPERT SYSTEM

Crowdacred

Overview

Crowdacred

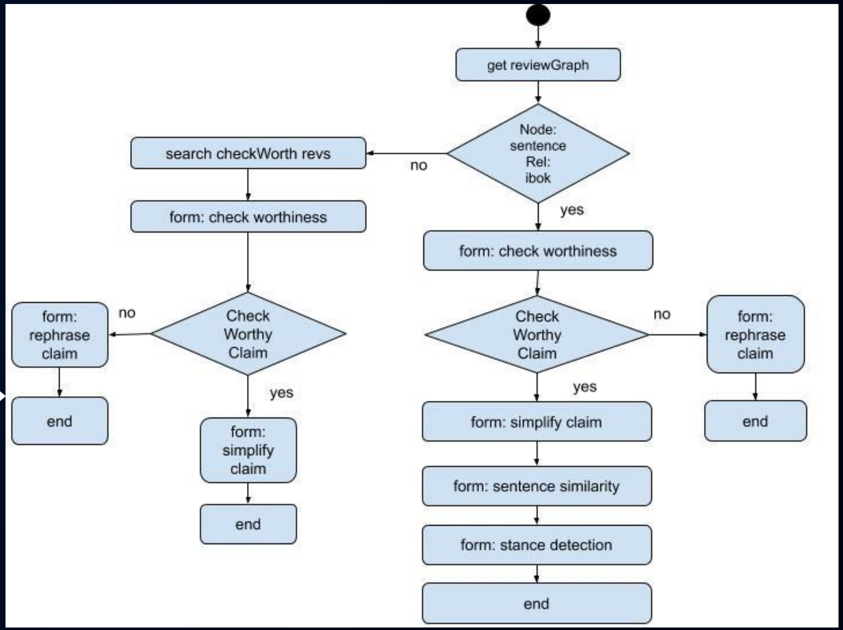
ReviewGraph



Dis-agree?

disagree

Detailed Feedback Wizard

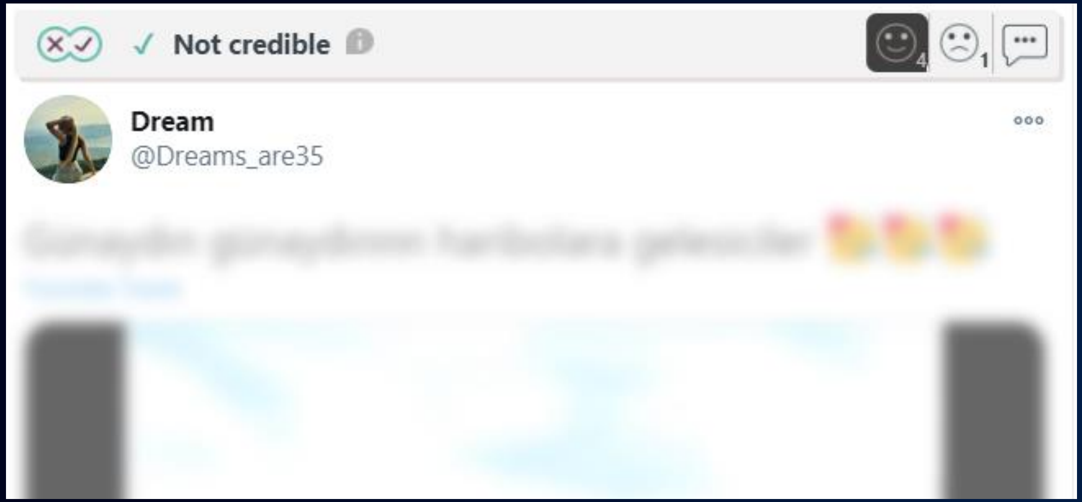


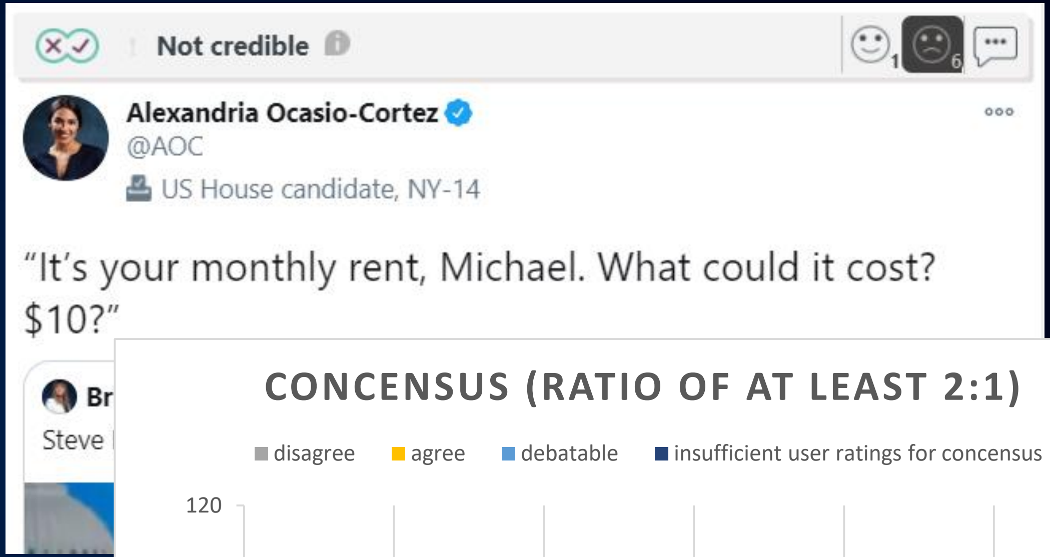
EXPERT SYSTEM

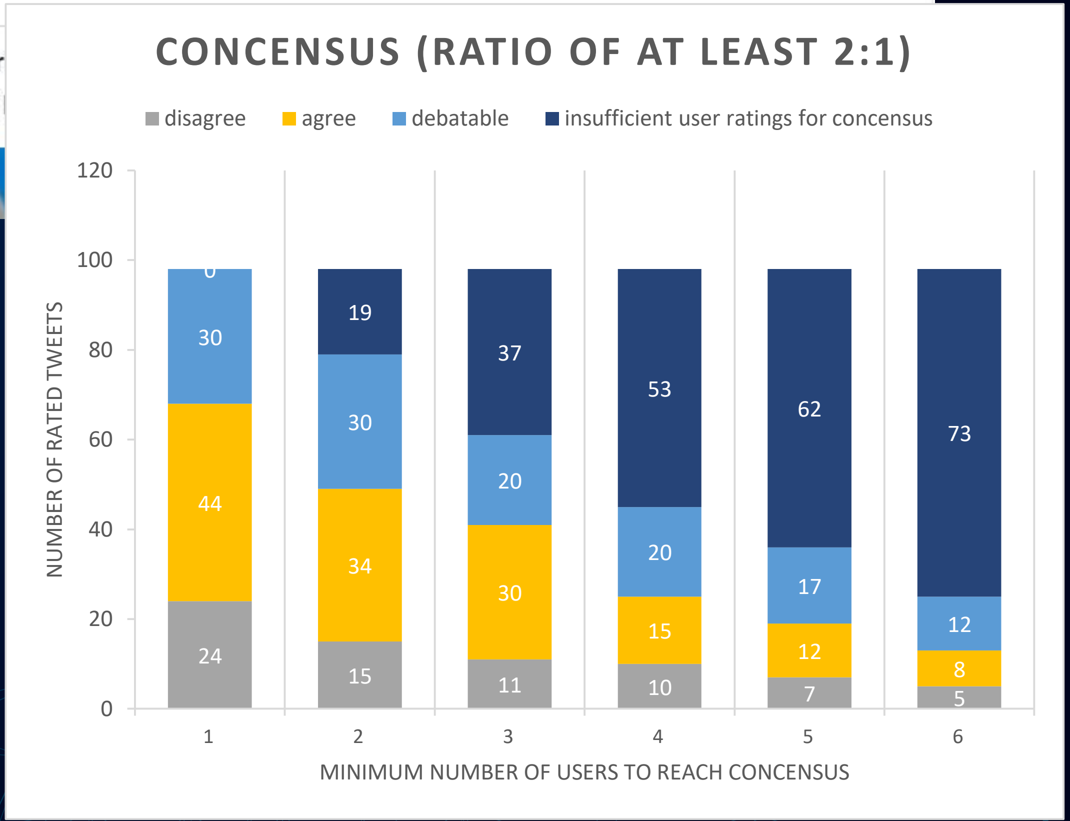
Crowdacred

Dis/agreement

Tweet + Label (explanation optional)







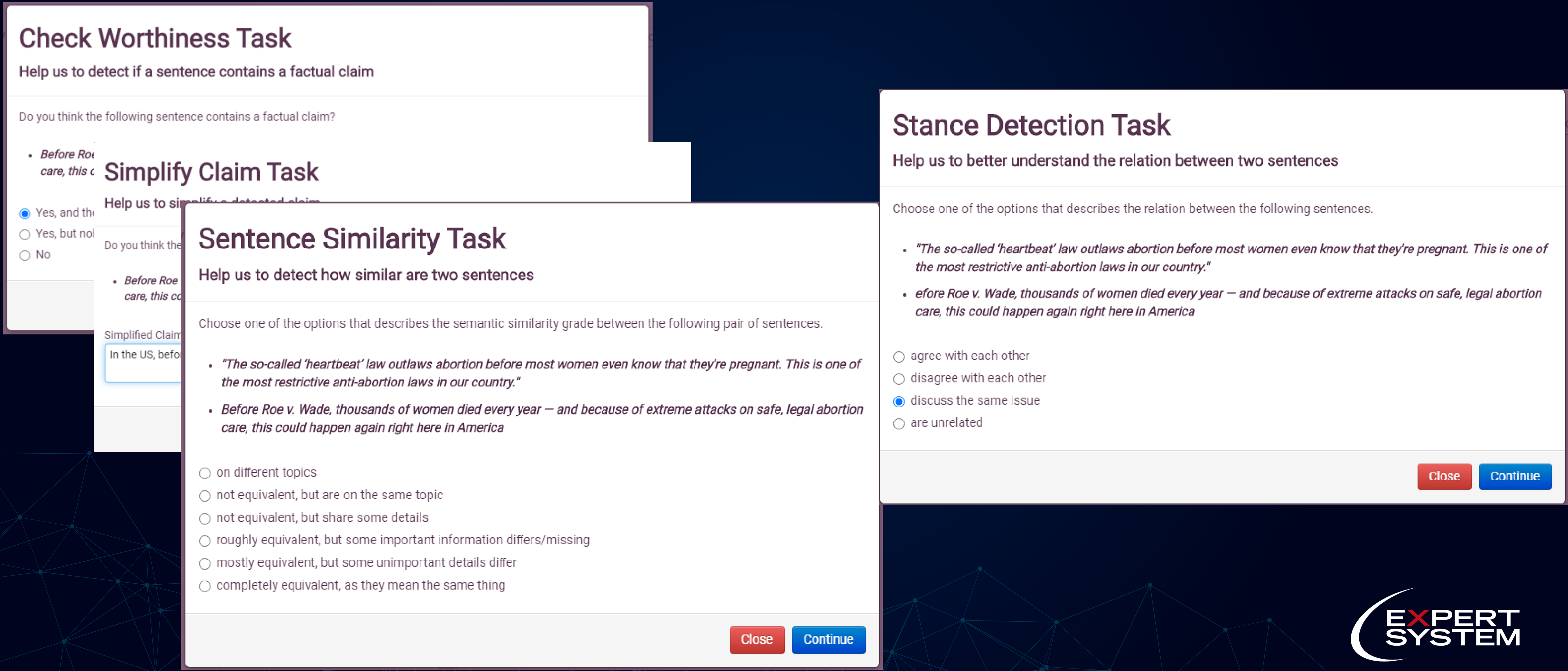
Can we reach consensus?

- At least n (dis)agreements for tweet/label
- At least d:a ratio

Currently collecting ratings for 400 tweets in coinform4550_train using Co-inform Browser Plugin

Crowdacred

Detailed feedback wizard



EXPERT SYSTEM

Conclusion and Future work

Dis/agreement phase seems to converge

Useful for identifying errors
What to do with “debatable” cases?
How many “agree” cases are lucky?

Detailed Feedback Wizard

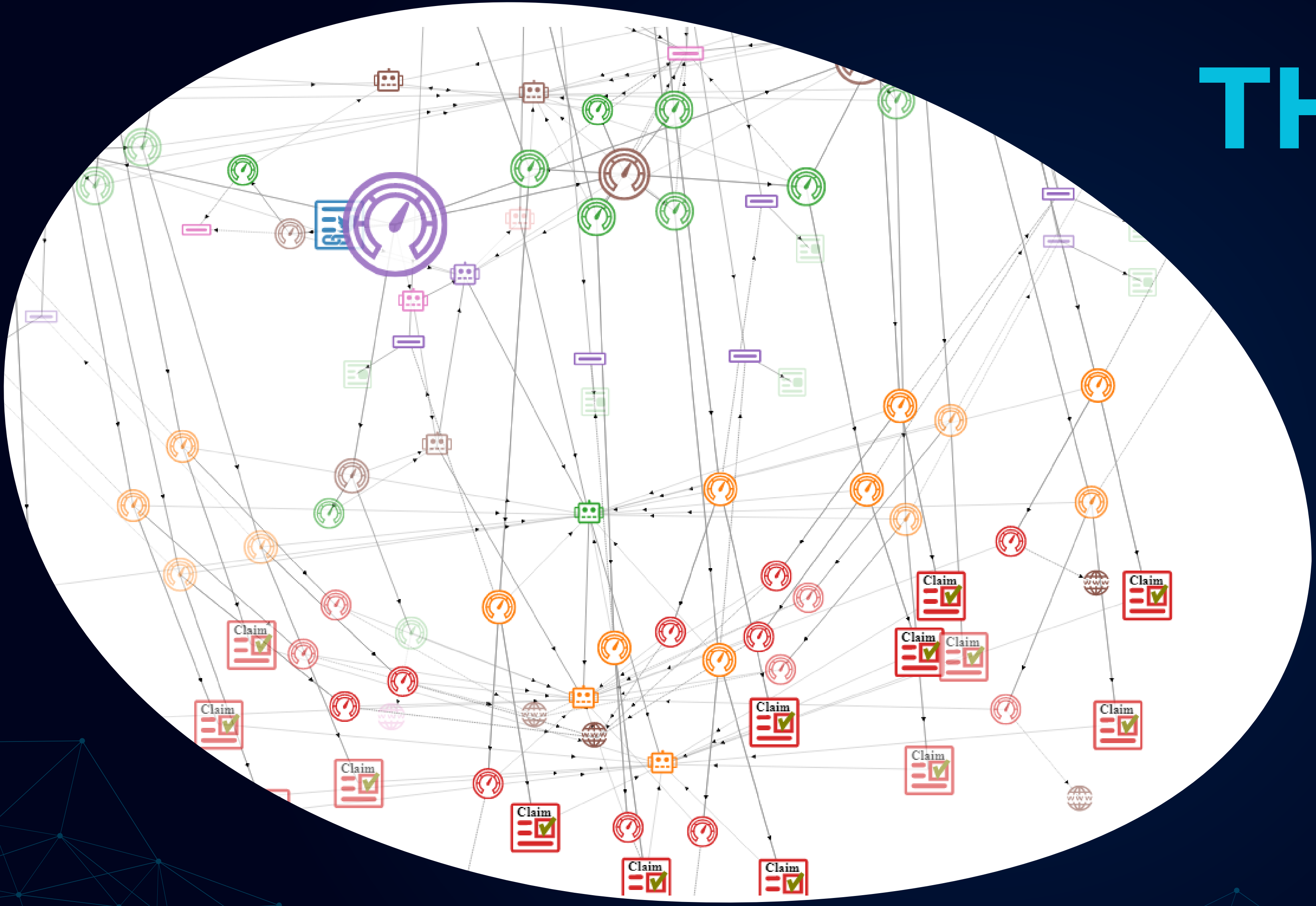
Now able to:

- Prune Review Graph to focus on relevant steps/sub reviews
- Generate simple tasks to get feedback and generate new datasamples

Open Questions

How to combine detailed feedback from different users?
How many new data samples do we need to improve RoBERTa models?
Will improvements generalise?

THANK YOU



Co-inform

EXPERT SYSTEM