COVID-Fact: Fact Extraction and Verification of Real-World Claims Concerning the COVID-19 Pandemic

Arkadiy Saakyan
a.saakyan@columbia.edu

Tuhin Chakrabarty
tuhin.chakr@cs.columbia.edu

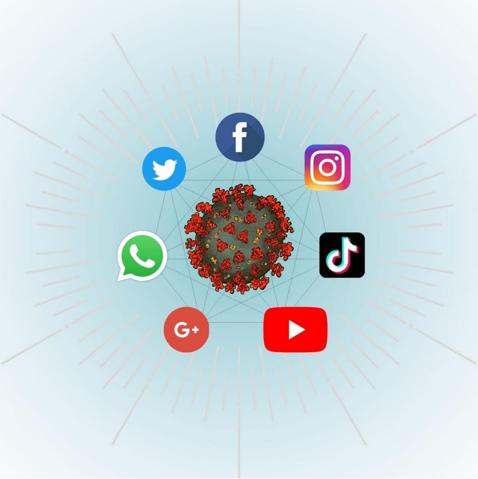
Smaranda Muresan
smara@columbia.edu

Natural Language Processing
Columbia University

INDIA'S HEALTHCARE WORKERS ARE BUSTING MISINFORMATION ON WHATSAPP

The backbone of India's rural healthcare system is now tasked with beating back COVID-19 myths, one message at a time

By Sanket Jain | Jun 17, 2021, 9:00am EDT



What Should an Ideal Fact-checking System Do?

1. Consider real-world claims Multi-FC (Augenstein et al., 2019)
2. Retrieve relevant documents not bound to a known document collection (i.e. Wikipedia) which can validate the claim Multi-FC (Augenstein et al., 2019)
3. Select evidence sentences that support or refute the claim
   FEVER (Thorne et al., 2018, 2019), SciFact (Wadden et al., 2020)
4. Predict claim veracity based on evidence
   FEVER (Thorne et al., 2018, 2019), SciFact (Wadden et al., 2020), Multi-FC (Augenstein et al., 2019)

- Current research addresses several of these tasks, but not all
- Refuted claims are typically generated via crowdsourcing

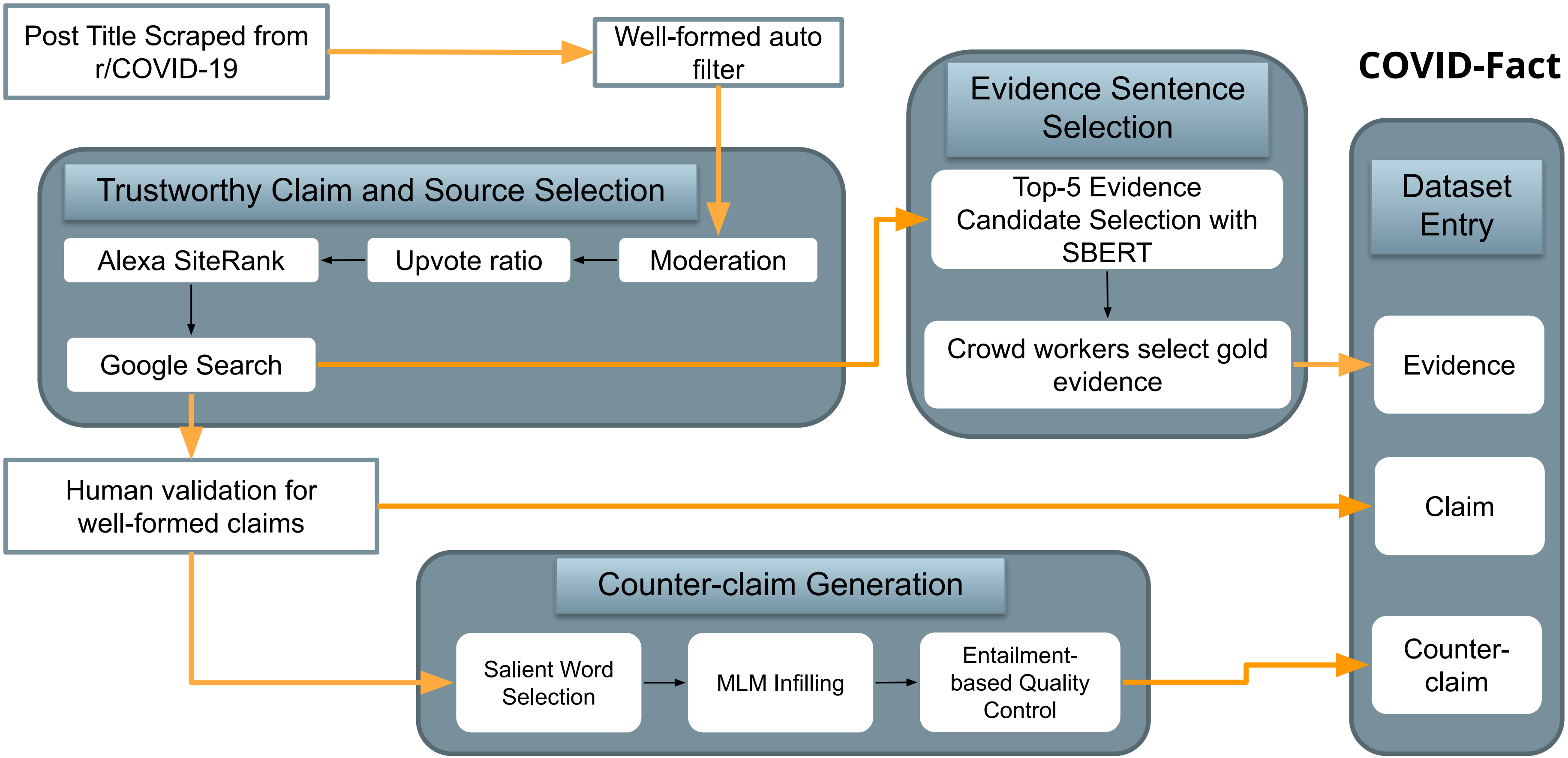
COVID-Fact: Our Contributions

- Automatic *real-world true claim* and *trustworthy* evidence document selection
- *Automatic generation of counter-claims* from true claims
- Evidence sentence selection using textual similarity and crowdsourcing
- Dataset of *4,086 real-world claims on the COVID-19* pandemic annotated with *sentence-level evidence*
    - Veracity prediction baseline models + zero-shot performance on SciFact

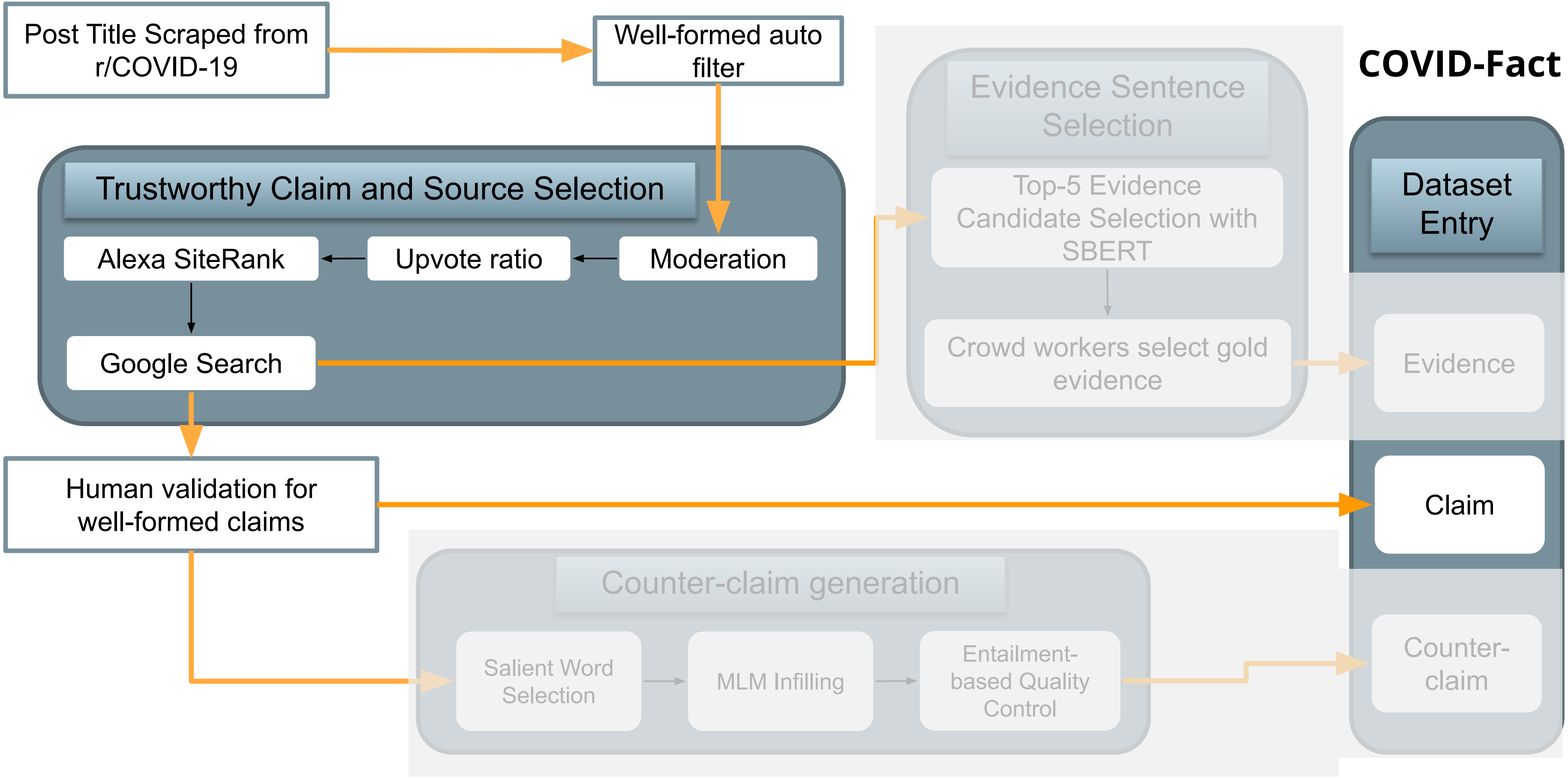
Original Claim	Closed environments **facilitate** secondary transmission of coronavirus disease 2019
Counter-Claim	Closed environments **prevent** secondary transmission of coronavirus disease 2019
Gold Document	`https://www.medrxiv.org/content/10.1101/2020.02.28.20029272v2`
Gold Evidence	It is plausible that closed environments contribute to secondary transmission of COVID-19 and promote superspreading events.

Original Claim and Generated Counter Claim **SUPPORTED** and **REFUTED** by same evidence

COVID-Fact Overview

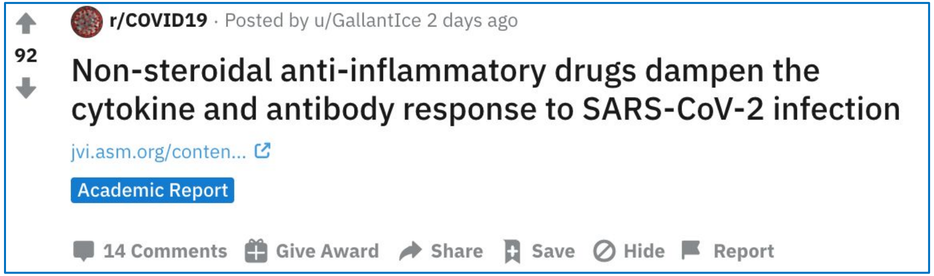


Trustworthy Claim and Source Selection

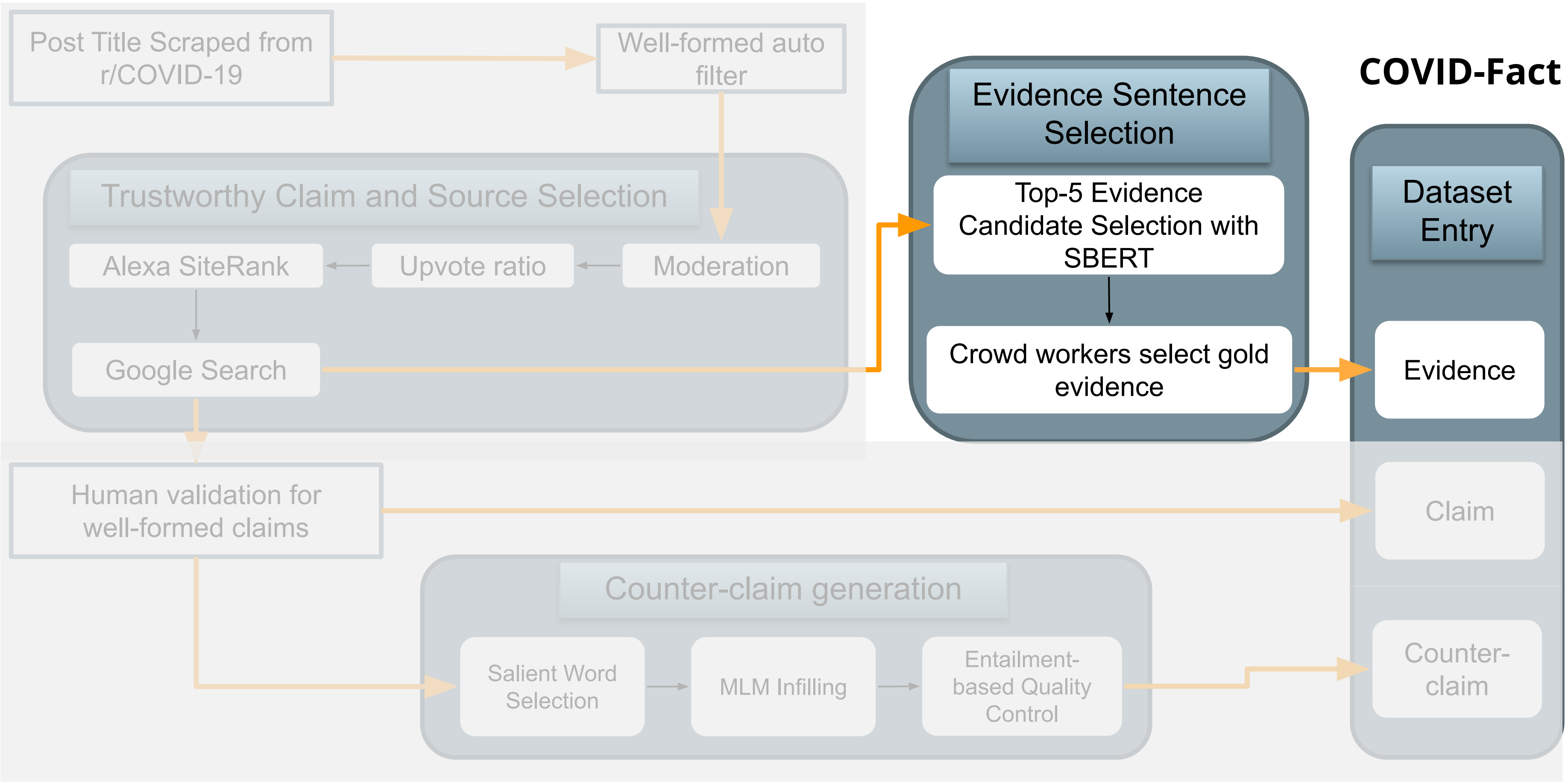


Trustworthy Claim and Source Document Selection

- Scraped titles of posts from the **r/COVID-19** subreddit
  - ***Real-world scientific claims***
  - ***Claims expressed in lay language***
- Filter for well-formed claims
- Filtering for trustworthiness:
  - Subreddit moderators
  - Upvote ratio > 0.7
  - Alexa SiteRank < 50,000
  - Google Search results



Evidence Sentence Selection



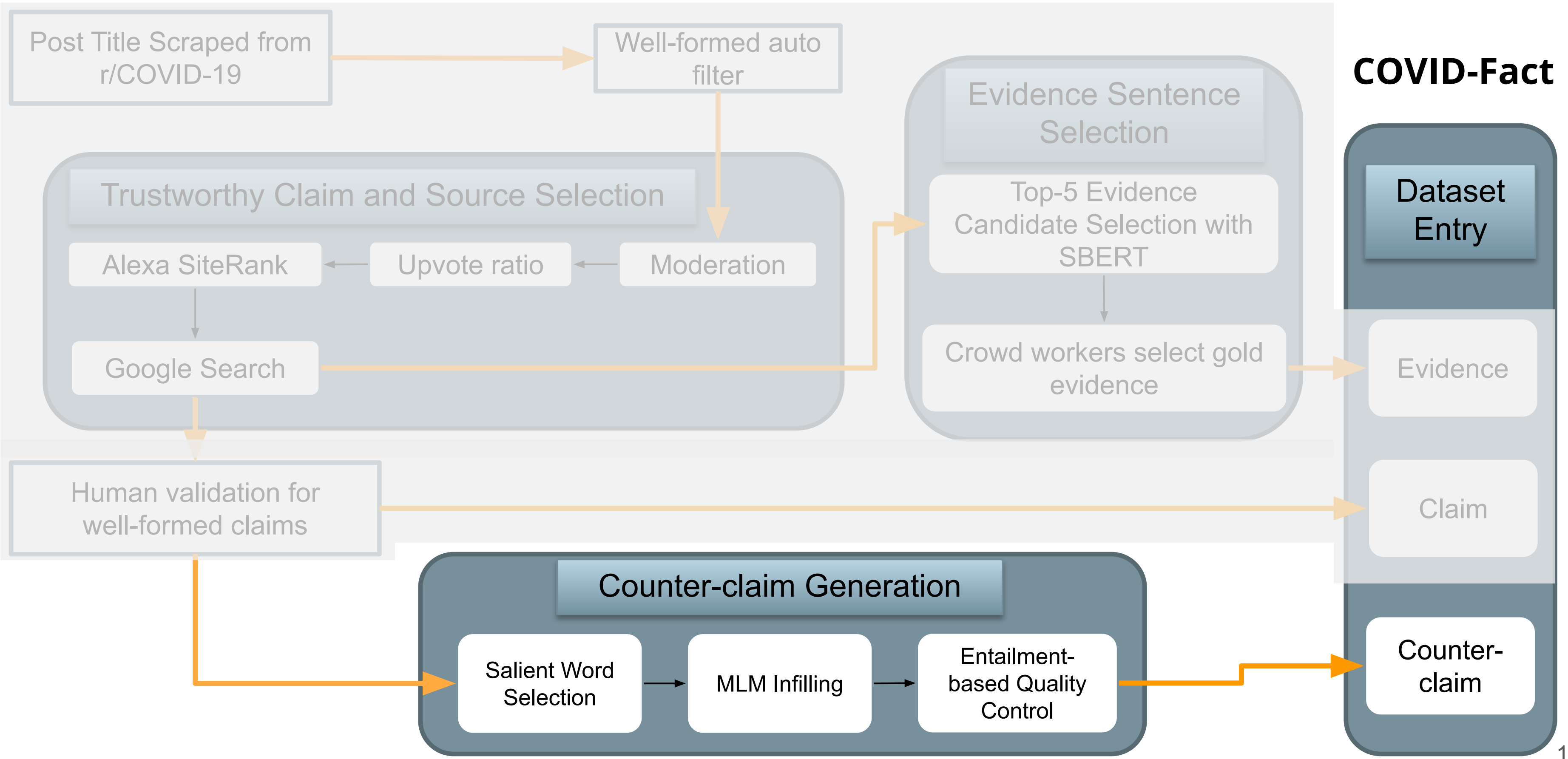
Evidence Sentence Selection

- Use cosine similarity on SBERT sentence embeddings (Reimers and Gurevych, 2019) to extract top five sentences most similar to the true claim from the top 5 Google Search result pages

- Amazon Mechanical Turk crowdworkers select which of these sentence constitute evidence for the claim (or select absent if no evidence)

- Only need this for *supported claims*, the corresponding refuted claims will have the same evidence

Automatic Counter-claim Generation



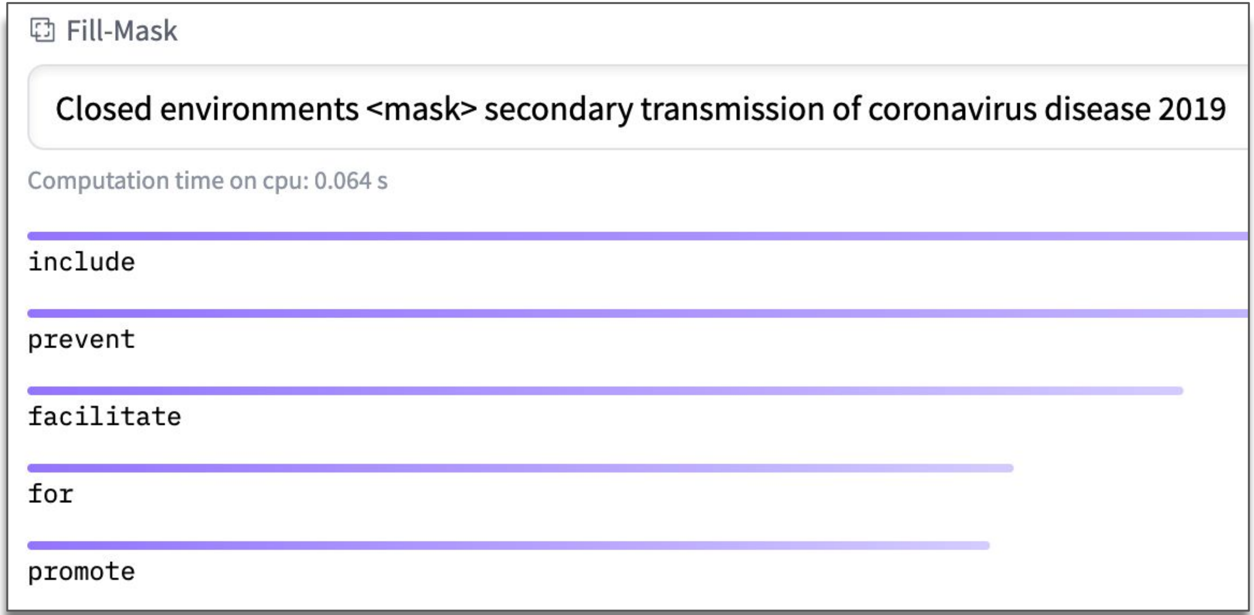
Automatic Counter-Claim Generation (1)

- ***Salient Word Selection***: given a *true* claim select salient words to replace
  - Fine-tune BERT to classify Supported vs Refuted claims in the SciFact dataset and use the model to extract attention scores to find salient words
  - Attention-based salience: 68% recall with human judgments

*"Closed environments **facilitate** secondary transmission of Coronavirus disease 2019"*

Automatic Counter-Claim Generation (2)

Masked Language Model Infilling: Use RoBERTa fine-tuned on CORD-19 to replace salient words with top-k candidates (ignores grammatically incorrect candidates)



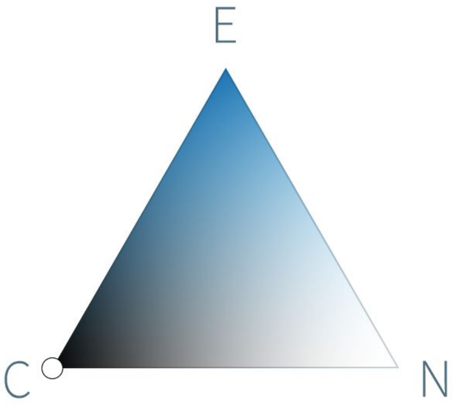
Automatic Counter-Claim Generation (3)

Entailment-based Quality Control: select top 3 claims that achieve a contradiction score above 0.9 using RoBERTa fine-tuned on MultiNLI

Premise: Closed environments **facilitate** secondary transmission of coronavirus disease 2019

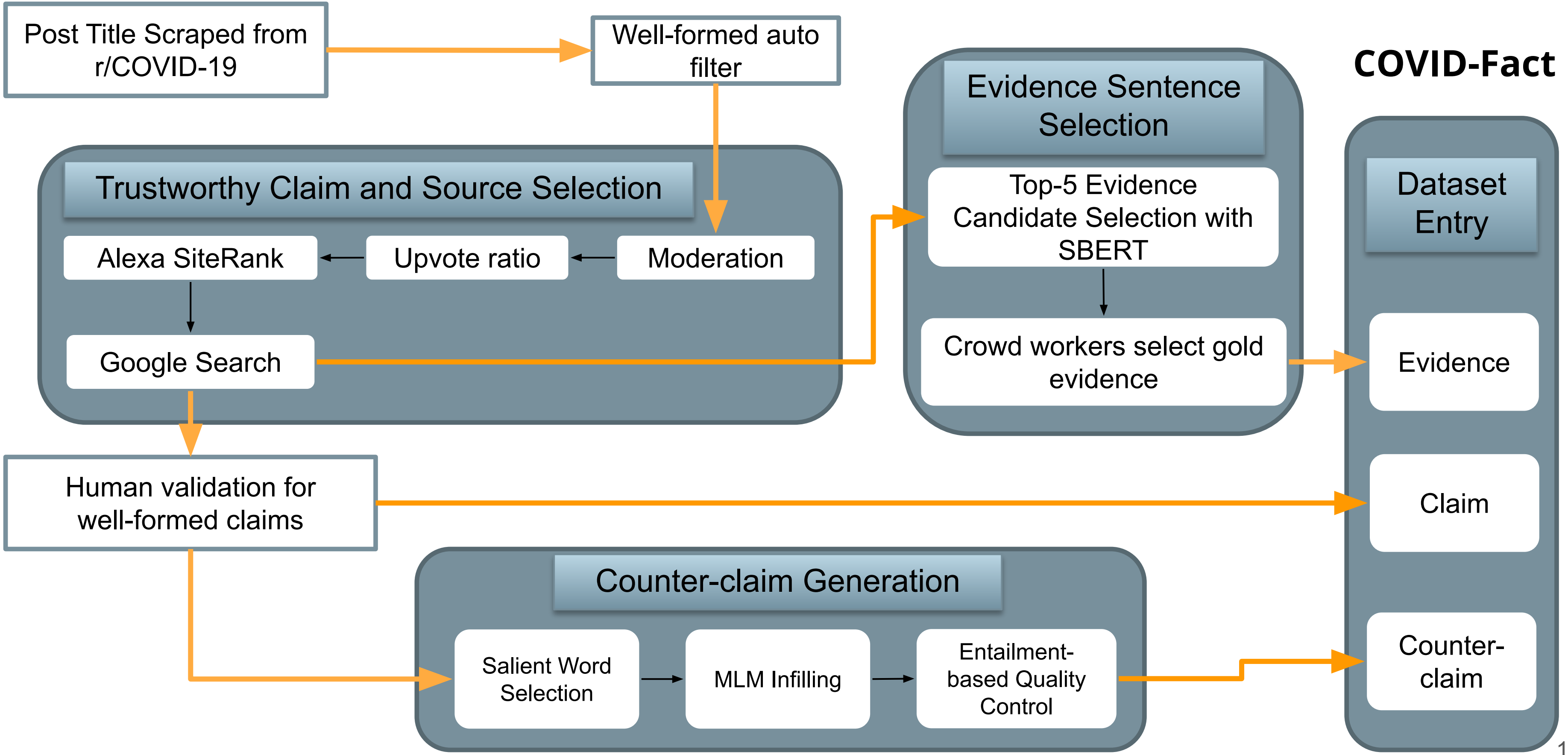
Hypothesis: Closed environments **prevent** secondary transmission of coronavirus disease 2019

Contradiction score by RoBERTa MNLI: 99.8%



Judgement	Probability
Entailment	0.1%
Contradiction	99.8%
Neutral	0.1%

Claim and Evidence Selection: Overview



COVID-Fact Task Formulation

- **Task**: Given a claim **c**, a system must retrieve a ***set of evidence sentences***, and determine a label **v** ∈ {SUPPORTED, REFUTED} based on this evidence.

- **Metric: COVID-FEVER Score (veracity prediction + evidence retrieval)**
  - *1 if it correctly predicts the veracity of the claim-evidence pair and if at least one of the predicted evidence matches the gold evidence selected by annotators (thus a stricter score than veracity prediction accuracy).*

Baseline pipeline for the COVID-Fact Task

- **Evidence Retrieval**:
  - Google search to identify five potential source documents by querying the claim
  - Select most similar sentences using cosine similarity between sentence embeddings of the claim and candidate sentences using SBERT

- **Veracity prediction:**
  - RoBERTa model.
  - Concatenate all evidence sentences in the evidence set and use it as input for a binary classification task.

Results: Veracity Prediction and COVID-FEVER

Veracity Prediction								COVID-FEVER
	Gold		Top 5		Top 1			Top 5
	Acc	F1	Acc	F1	Acc	F1		Score
MNLI (Williams et al., 2018)	61.3	64.2	53.1	51.5	65.4	60.6		35.1
SciFact (Wadden et al., 2020)	56.9	57.0	53.7	54.0	54.3	54.0		36.9
FEVER (Thorne et al., 2018)	48.3	47.0	46.2	45.0	48.6	48.0		35.4
COVID-Fact	**83.5**	**82.0**	**84.7**	**83.0**	**83.2**	**81.0**		**43.3**
SciFact + COVID-Fact	82.2	81.0	83.0	82.0	80.2	79.0		43.0
FEVER + COVID-Fact	74.8	70.0	78.2	73.0	73.3	68.0		35.4
COVID-Fact (Claim only)	67.5	40.0	-	-	-	-		-

- Given gold evidence: fine-tuning on COVID-Fact led to performance improvement of **25** F1-score and **35** F1-score compared to training solely on SciFact and FEVER, respectively.
- Our baseline pipeline achieves a COVID-FEVER score of 43.3 using Top-5 evidence sentences
- Adding the FEVER and SciFact datasets deteriorates the results.

Usefulness of COVID-Fact for Zero-Shot Scientific Fact-checking

- We train models on COVID-Fact claims and gold evidence and evaluate the veracity performance on the ***SciFact*** dev set in a zero-shot setting.

- ***SciFact*** only contains scientific claims, model trained only on SciFact does not generalize well to COVID-Fact, which also contains non-scientific claims. COVID-Fact, on the other hand, contains enough scientific claims so that the model generalizes well to ***SciFact***.

Train Setting	Acc	F1
COVID-Fact	80.8	80.0
Sci-Fact	83.7	83.0

Error Analysis

- Cause and Effect
- Commonsense Knowledge
- Scientific Background

C1	SARS-CoV-2 is **not detectable** in the vaginal fluid of **women** with severe COVID-19 infection
EV1	**All 10 patients** were tested for SARS-CoV-2 in vaginal fluid,and all samples tested **negative** for the virus.
C2	Baricitinib **restrains the immune dysregulation** in COVID-19 patients
EV2	Here, we provide evidences on the efficacy of Baricitini, a JAK1/JAK2 inhibitor, in **correcting the immune abnormalities** observed in patients hospitalized with COVID-19.

Conclusion

- Dataset of *4,086 real-world claims on the COVID-19* pandemic annotated with *sentence-level evidence*
- Automatic *real-world true claim* and *trustworthy* evidence document selection
- *Automatic generation of counter-claims* from true claims
- Evidence sentence selection using textual similarity and crowdsourcing

Thank you!

a.saakyan@columbia.edu
tuhin.chakr@cs.columbia.edu
smara@columbia.edu