A FACT EXTRACTION AND VERIFICATION FRAMEWORK FOR SOCIAL MEDIA POSTS ON COVID-19

Orkun Temiz, **Tuğba Taşkaya Temizel**

INFORMATICS INSTITUTE

2 April 2023

Problem Definition

- The tremendous amount of information and user posts increased the necessity of fact-checking and its spread, directly threats individuals, organizations and public health.
- In COVID-19 pandemic, misinformation disseminates faster than the virus its prevention has become one of the main concerns.
- Fact-checking organizations rely on independent journalists such as Snopes, and Politifact have increased their activities.
- Failed to timely respond to COVID-19 misinformation as manual preparation of each claim’s response takes a significant amount of time.
- There is a strong need for an automatic, almost real-time fact-checking solution to detect misinformation as well as to verify given information.

Research Questions

- Can we develop a method for fact-checking and verifying user posts against published peer-reviewed articles? If so, how?
  - How can we map an informal medical claim to formal medical articles on social media?
  - How can we develop an evidence-based fact-checking method without direct supervision? How does it perform compared to the state-of-art supervised models?
  - Can we improve medical document retrieval performance by using MeSH* (Medical Subject Headings) tree structure?

*The **Medical Subject Headings** (MeSH) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information

Contributions of Study

- Mapping informal user posts to scientific medical articles and fact check these posts against the related evidence found in these articles.
- Providing a framework utilizing the zero-shot capabilities of the existing models to fact-check user posts, including medical claims without explicit supervision.
- Providing evidence-based fact-checking linked to a medical article/sources empowers users to make up their minds considering explicit references related to the claim

Data Sources

- For claims:
  - CoAid Dataset (with Tweet ids)
    - User Claims
    - News
  - Comprises around 4,250 news, 300,000 related comments, and 900 social media posts.
  - Manually labeled ground truth claims were used as search queries to automatically retrieve related tweets and label them.
  - In the experiments, the tweets including solely user posts (henceforward called "*Claims*") or the tweets comprising "news titles" (henceforward called "*News*") are used.
  - Example:
    - Tweet: "Look at this and please share"
    - Title of the news: "New flu drug drives drug resistance in influenza viruses".

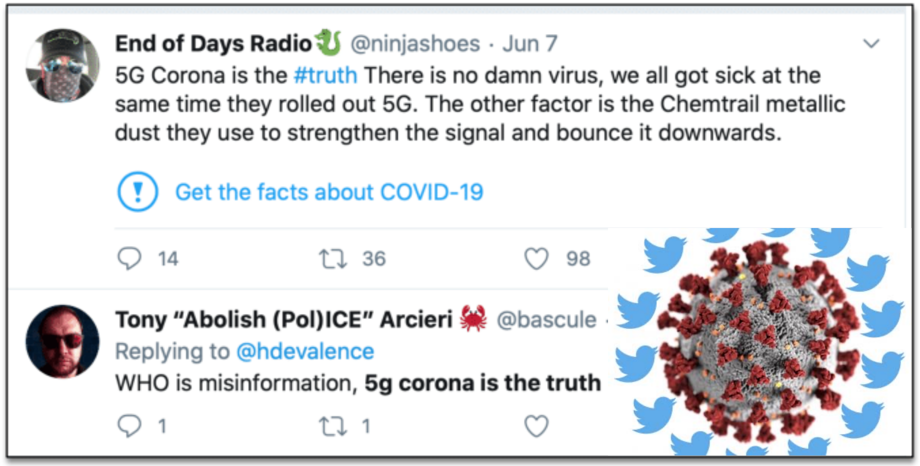
Data Sources

- COVID-19 Rumors Dataset (Raw User Posts)
  - Includes manually labeled 4,129 rumors from news and 2,705 rumors from user posts in Twitter with sentiment and stance labels.
  - The true status of the rumors was manually retrieved from fact-checking websites.
  - In term of content, it is mixed (having fewer medical claims compared to CoAID).

Data Sources

For evidences:

- COVID-19 Open Research Article Repository from Allen Institute for AI (AI2)
- PubMed
- MedrXiv



End of Days Radio @ninjashoes · Jun 7
5G Corona is the #truth There is no damn virus, we all got sick at the same time they rolled out 5G. The other factor is the Chemtrail metallic dust they use to strengthen the signal and bounce it downwards.

Get the facts about COVID-19

14 36 98

Tony "Abolish (Pol)ICE" Arcieri @bascule ·
Replying to @hdevalence
WHO is misinformation, **5g corona is the truth**

1 1

World Health Organization

PubMed Central

PubMed

bioRxiv

medRxiv

arXiv.org

The Gap with Related Works

1) Being able to react to new emerging claims

2) Being able to retrieve relevant documents from a regularly updated document collection such as MEDLINE, PubMed, MedriXv

3) Selecting textual evidence sentences that can support or refute the claim

4) Being able to establish a link between informal and formal texts to relate claims present in user posts with the evidence obtained from the scientific articles

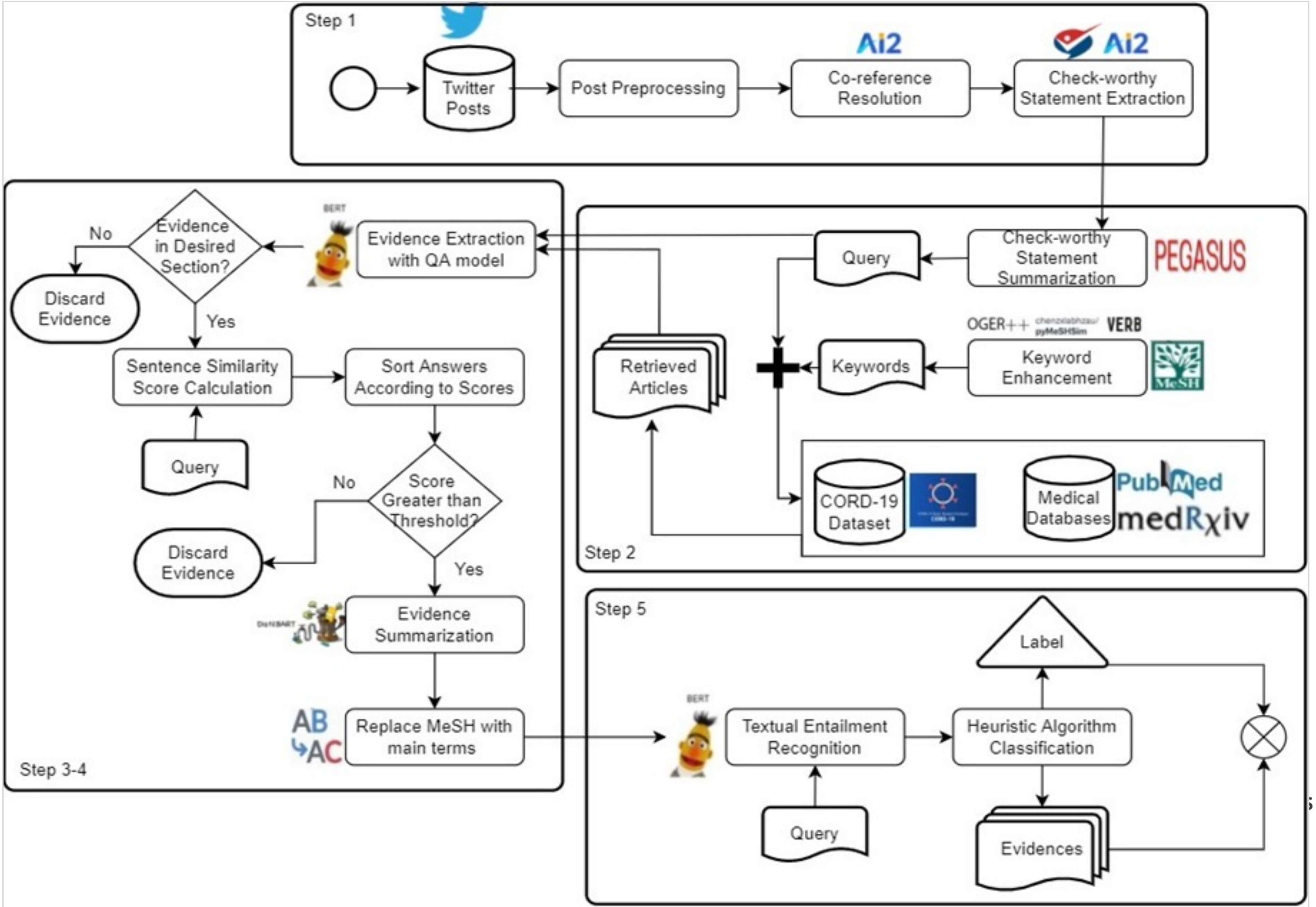
5) Being able to predict the claim's veracity based on the evidence collection.

Multi-FC [1]
Not evidence based

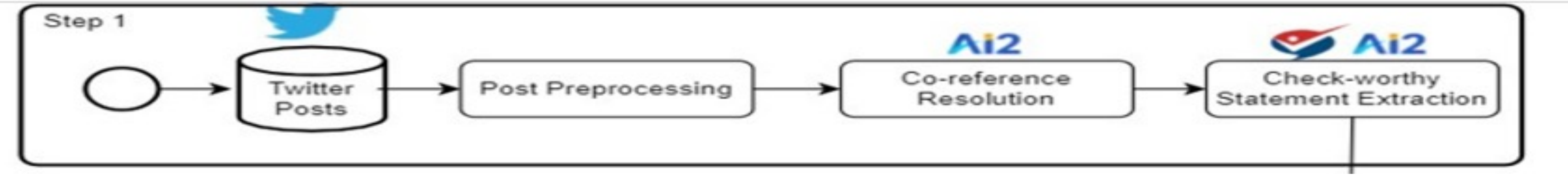
FEVER [2], SciFact
Static Collection
[3]

[1] Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing
[2] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A Largescale Dataset for Fact Extraction and VERification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 809–819. https://doi.org/10.18653/v1/N18- 1074 [3] Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

Proposed Framework



Framework: Step 1

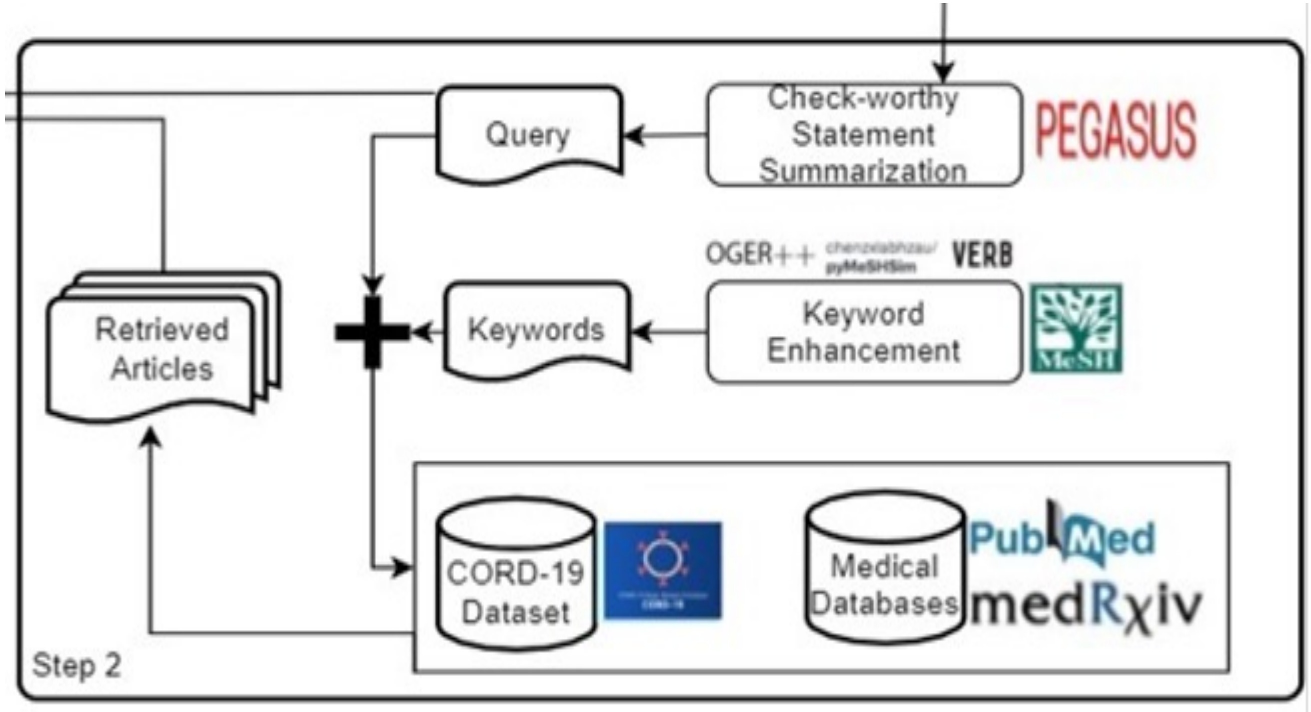


- **Preprocessing**
  - Strip out from special characters and irrelevant text e.g., hashtags, URLs, emojis, mentions, images etc.
  - Hashtag words are not removed directly to preserve the meaning of the posts. Hashtag words were further processed by constituency parser to differentiate hashtag chunks from hashtag words in a sentence.
- **Claim Extraction**
  - Identify sentence and noun phrase structures and discard hashtag chunks at the end of the sentence (AllenNLP Constituency Parser)
  - Identify interrogative sentences (WH words e.g. Who, What and Auxiliary words e.g Can you, Do you)
  - *Co-reference Resolution (SpanBERT)

*"*Can regularly rinsing nose with saline help prevent infection with the new coronavirus? No. There is no evidence that this protected people from infection with new coronavirus.*"

  - Determine user posts includes check-worthy sentence (ClaimBuster API)

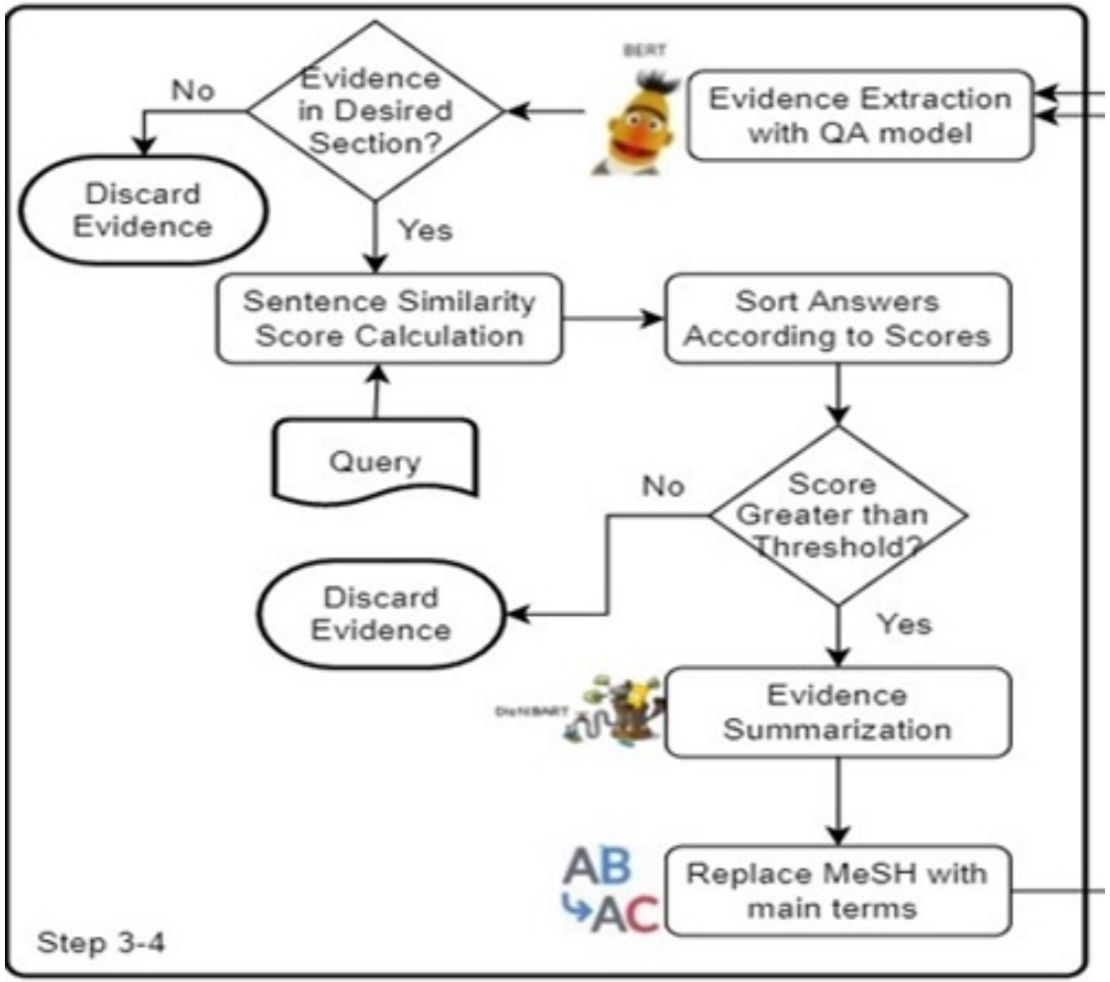
Framework: Step 2



- **Document Retrieval**
  - Anserini indexing on CORD-19 dataset
  - CORD-19, PubMed, medRxiv
  - Retrieve documents using (Keywords + Query)
  - Consider the articles published before the post’s publication date to prevent any data leakage and bias
- **Keyword Extraction & Enhancement**
  - Retrieve Medical Keywords (SciBERT)
  - Tokenization and Stemming
  - Medical Keyword Enhancement and Verification using Qualifier, Descriptor Terms and Supplementary Concept Record Terms for the keyword via MeSH Tree Structure (National Library of Medicine’s MeSH) e.g. 2019-nCov for COVID-19
  - Further enhancement (OGER++ and PyMeshSim)
  - Add verbs that directly dependent on medical keywords.

Framework: Step 3-4

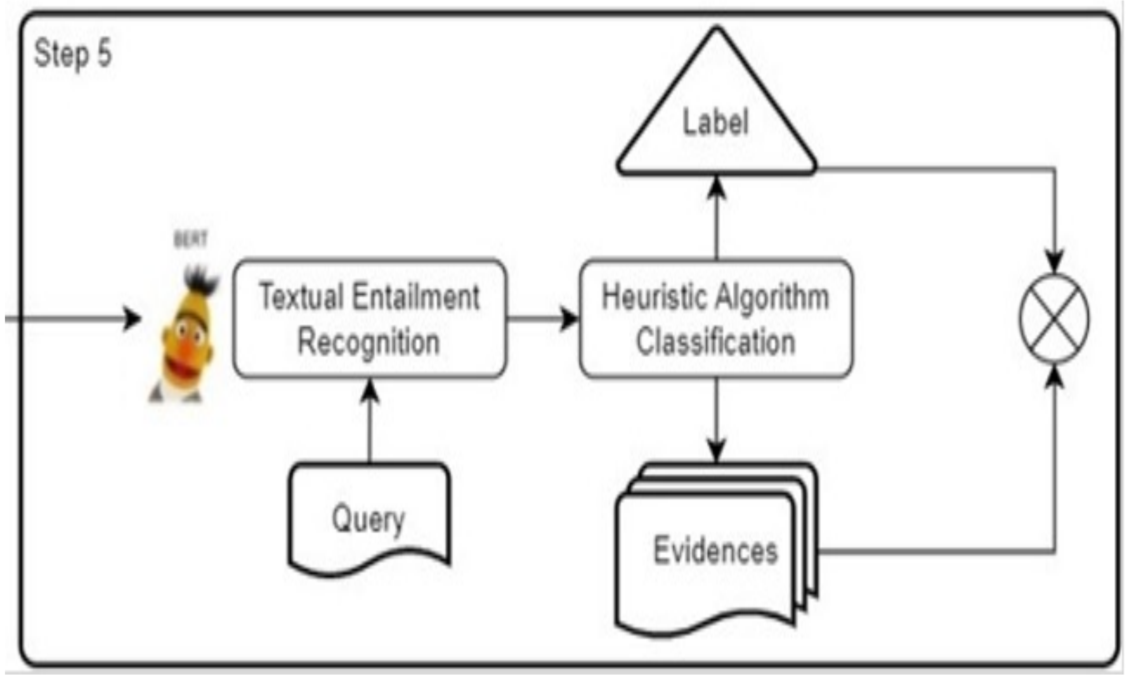
- **Evidence Selection**
  - Question Answering Model (BERT)
    - As a QA model, BIOBERT, based on BERT architecture, trained in the biomedical corpora is used.
  - Evidence Section Check
    - Only the abstract, introduction, discussion, result, or conclusion sections of an article are considered.
  - Sentence Similarity (Universal Sentence Encoder)
  - Evidence Summarization & Simplification (DistilBART)



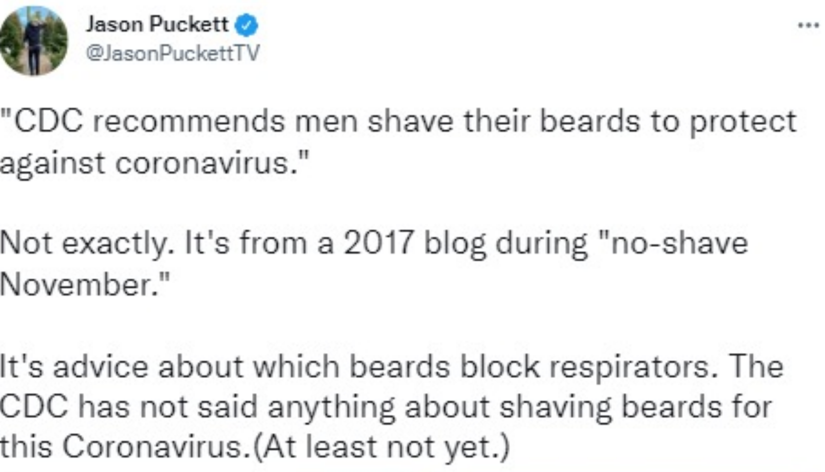
BERT
Evidence Extraction with QA model
Evidence in Desired Section?
No
Discard Evidence
Yes
Sentence Similarity Score Calculation
Query
Sort Answers According to Scores
Score Greater than Threshold?
No
Discard Evidence
Yes
DistilBART
Evidence Summarization
AB ↳AC
Replace MeSH with main terms
Step 3-4

Framework: Step 5

- **Textual Entailment**
  - Summarize Evidence (PEGASUS)
- **Heuristic Verdict Assignment**
  - Entailment scores calculated between the hypothesis and premise
  - Scores are used to label posts as ‘True’, ‘False’ or ‘Not Enough Info’


Step 5
Label
BERT
Textual Entailment Recognition
Heuristic Algorithm Classification
Query
Evidences

Example Tweet and News

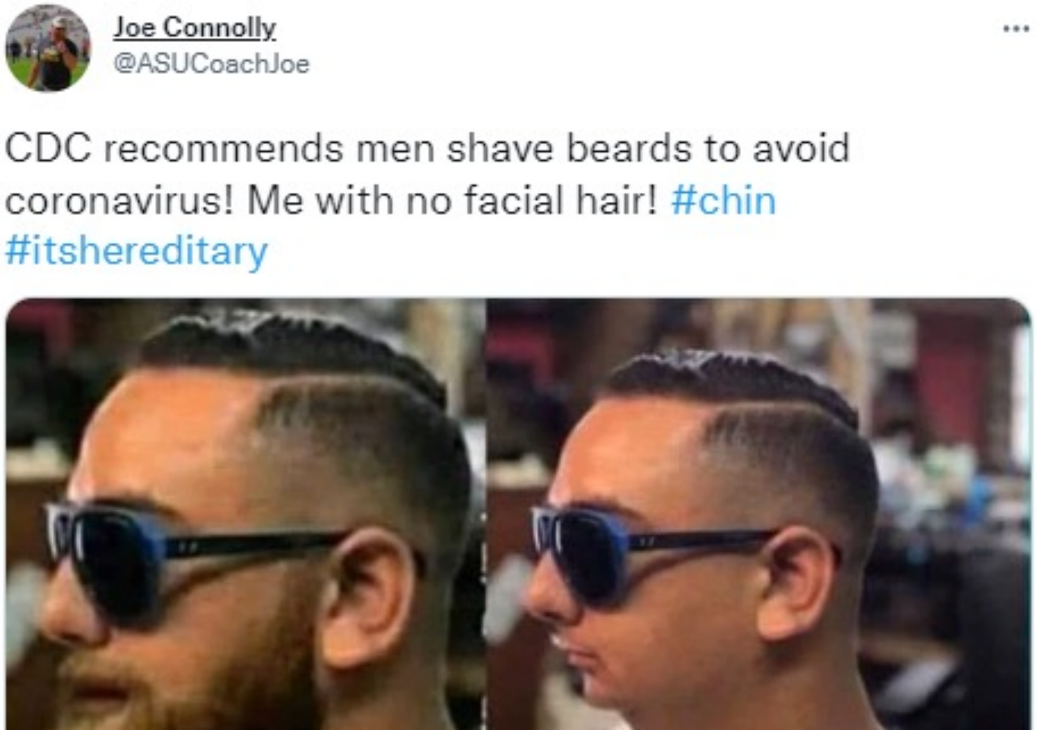


Jason Puckett
@JasonPuckettTV

"CDC recommends men shave their beards to protect against coronavirus."

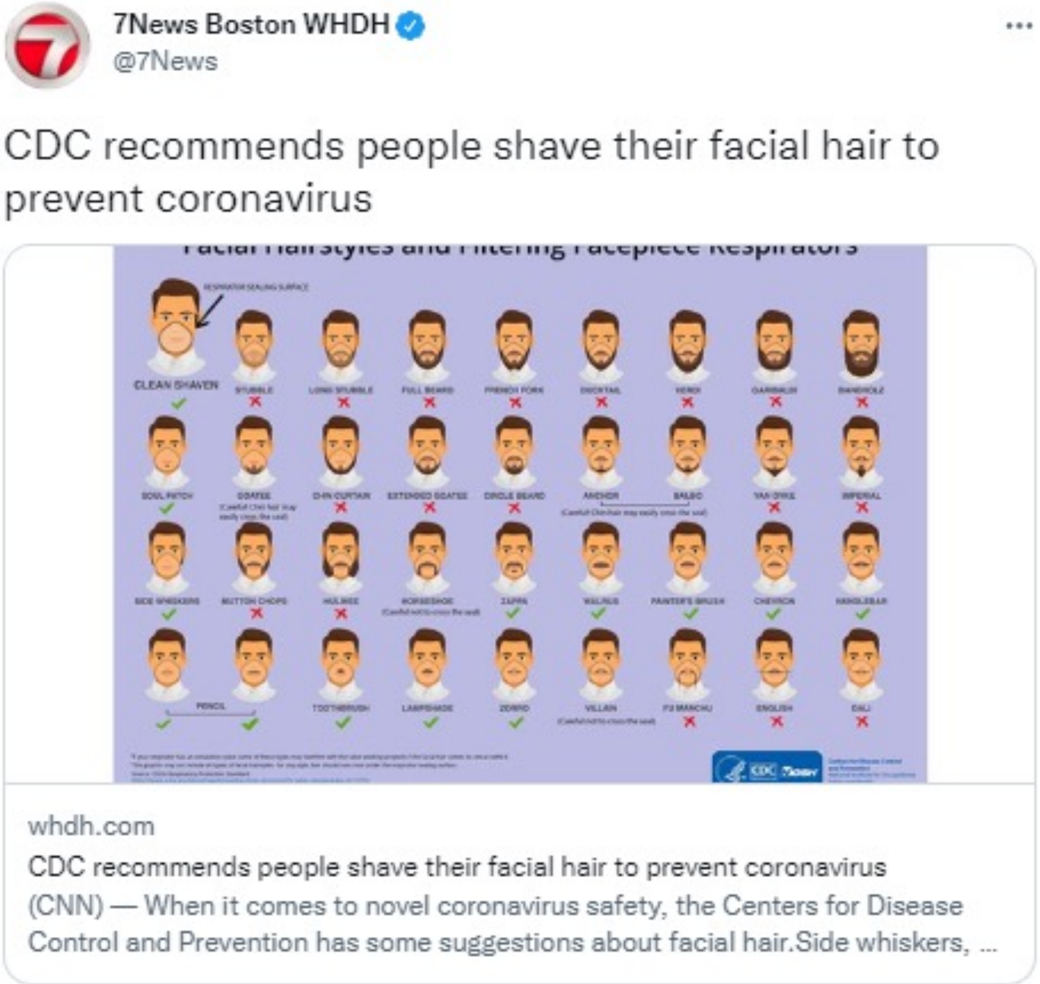
Not exactly. It's from a 2017 blog during "no-shave November."

It's advice about which beards block respirators. The CDC has not said anything about shaving beards for this Coronavirus.(At least not yet.)



Joe Connolly
@ASUCoachJoe

CDC recommends men shave beards to avoid coronavirus! Me with no facial hair! #chin #itshereditary

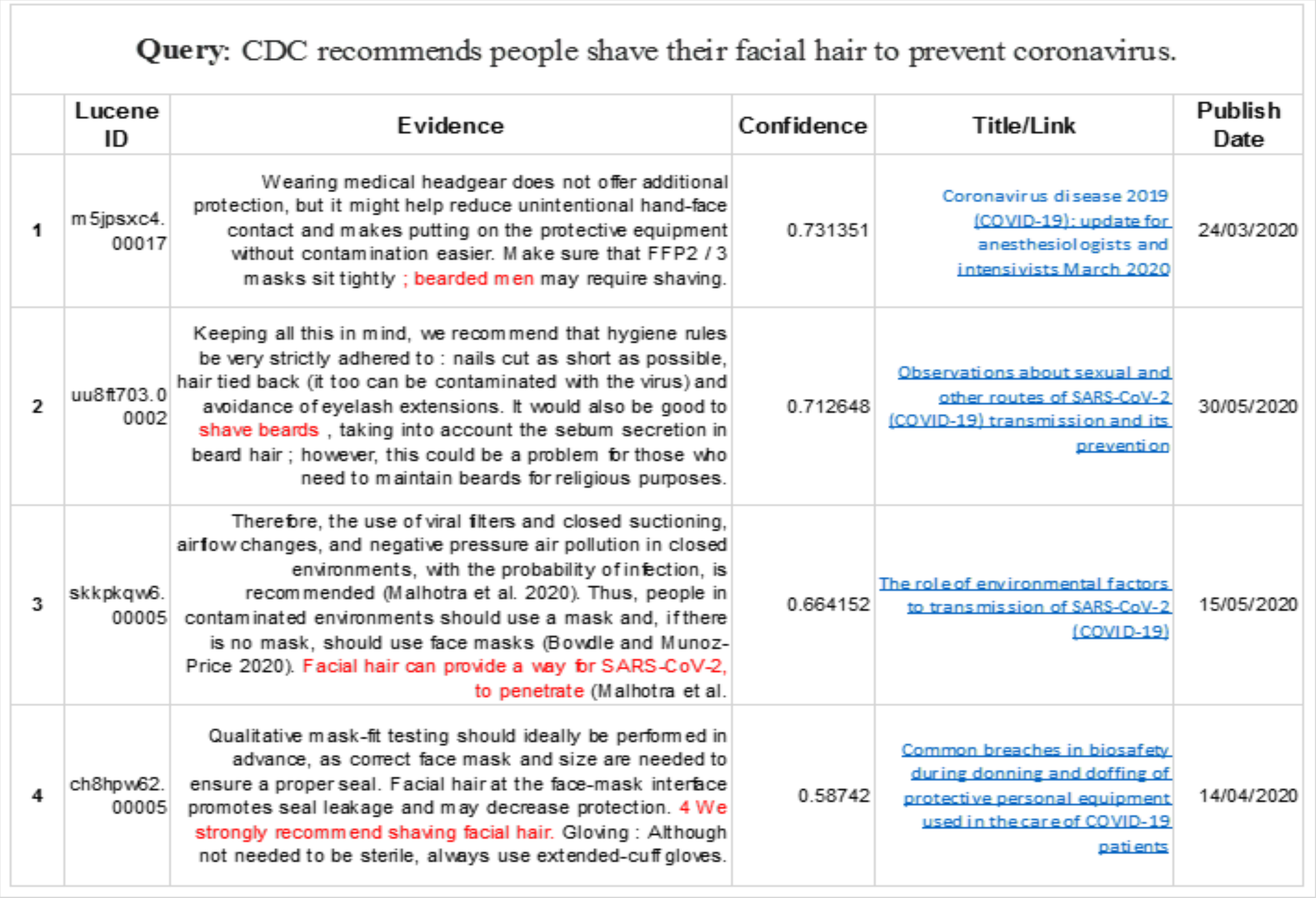


7News Boston WHDH
@7News

CDC recommends people shave their facial hair to prevent coronavirus

whdh.com
CDC recommends people shave their facial hair to prevent coronavirus
(CNN) — When it comes to novel coronavirus safety, the Centers for Disease Control and Prevention has some suggestions about facial hair.Side whiskers, ...

Example Outcome After Evidence Retrieval



Query: CDC recommends people shave their facial hair to prevent coronavirus.

	Lucene ID	Evidence	Confidence	Title/Link	Publish Date
1	m5jpsxc4.00017	Wearing medical headgear does not offer additional protection, but it might help reduce unintentional hand-face contact and makes putting on the protective equipment without contamination easier. Make sure that FFP2 / 3 masks sit tightly ; bearded men may require shaving.	0.731351	Coronavirus disease 2019 (COVID-19): update for anesthesiologists and intensivists March 2020	24/03/2020
2	uu8ft703.00002	Keeping all this in mind, we recommend that hygiene rules be very strictly adhered to : nails cut as short as possible, hair tied back (it too can be contaminated with the virus) and avoidance of eyelash extensions. It would also be good to shave beards , taking into account the sebum secretion in beard hair ; however, this could be a problem for those who need to maintain beards for religious purposes.	0.712648	Observations about sexual and other routes of SARS-CoV-2 (COVID-19) transmission and its prevention	30/05/2020
3	skkpkqw6.00005	Therefore, the use of viral filters and closed suctioning, airflow changes, and negative pressure air pollution in closed environments, with the probability of infection, is recommended (Malhotra et al. 2020). Thus, people in contaminated environments should use a mask and, if there is no mask, should use face masks (Bowdle and Munoz-Price 2020). Facial hair can provide a way for SARS-CoV-2, to penetrate (Malhotra et al.	0.664152	The role of environmental factors to transmission of SARS-CoV-2 (COVID-19)	15/05/2020
4	ch8hpw62.00005	Qualitative mask-fit testing should ideally be performed in advance, as correct face mask and size are needed to ensure a proper seal. Facial hair at the face-mask interface promotes seal leakage and may decrease protection. 4 We strongly recommend shaving facial hair. Gloving : Although not needed to be sterile, always use extended-cuff gloves.	0.58742	Common breaches in biosafety during donning and doffing of protective personal equipment used in the care of COVID-19 patients	14/04/2020

Retrieved answers are not always consistent…

- In some cases mixed results are encountered, where only some of the articles support the claim in the query while the remaining support the opposite.
- This phenomenon has been frequently observed in the COVID-19 pandemic since it was a novel coronavirus, and the preventive measures/treatments concerning the virus have constantly been changing over time.

WHO had previously said "*There was no need for the members of the public wearing a mask unless they were sick or around people with the coronavirus*."

Then 8.07.2020 – The WHO has changed its stance on wearing facemasks during the COVID-19 pandemic and said, "*WHO advises that governments should encourage the public to wear masks where there is widespread transmission in crowded environments and public transportation*."

Experiments

First Evaluation Scheme

- Uses an out-of-time sampling approach:
  - Tweets are sorted according to the dates of posts.
  - Then, we incrementally split the dataset timewise.
- Results:
  - Similar tweets may have been posted at different times, hence appearing in both training and testing datasets (data-leakage problem)

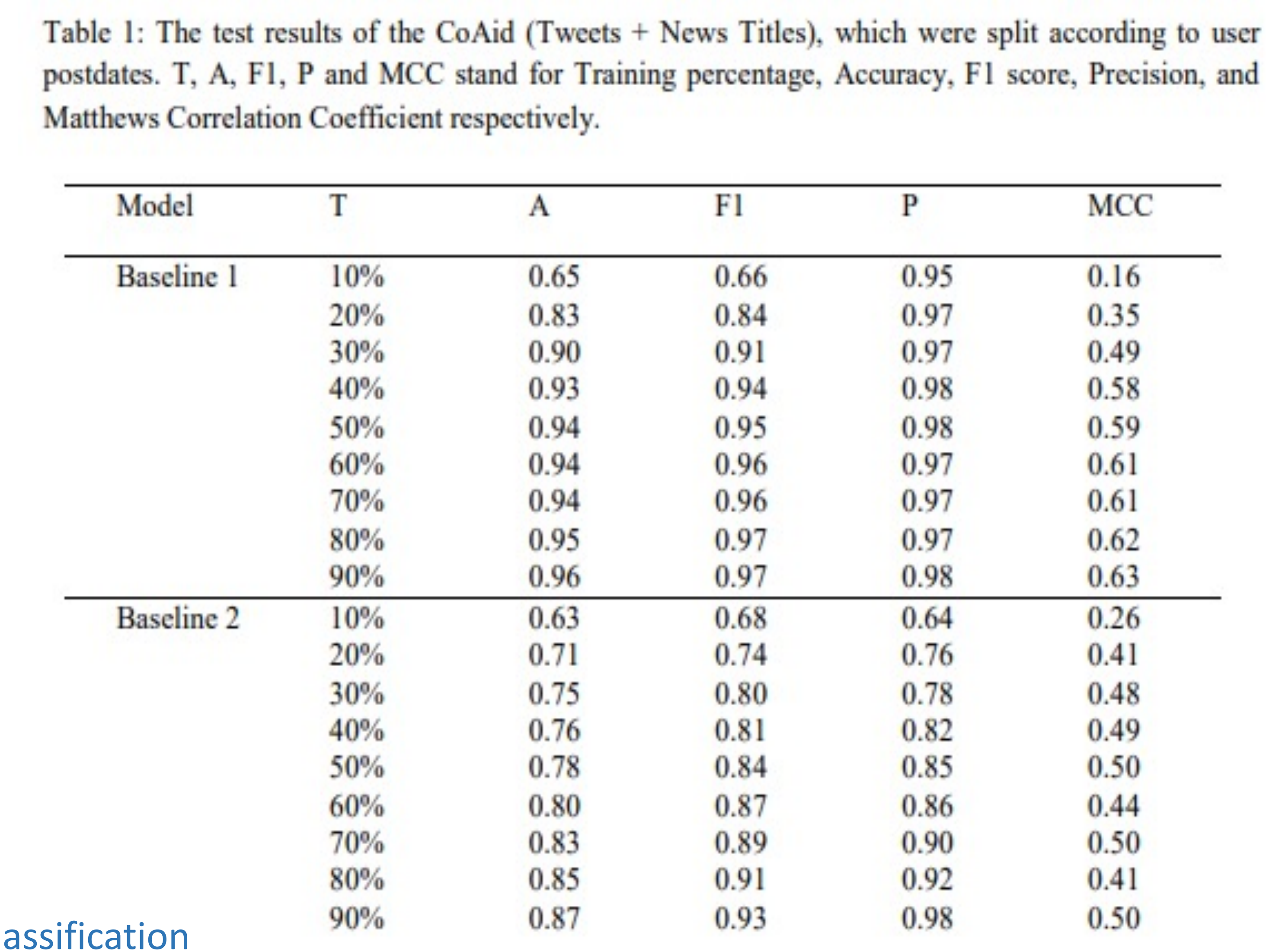


Table 1: The test results of the CoAid (Tweets + News Titles), which were split according to user postdates. T, A, F1, P and MCC stand for Training percentage, Accuracy, F1 score, Precision, and Matthews Correlation Coefficient respectively.

Model	T	A	F1	P	MCC
Baseline 1	10%	0.65	0.66	0.95	0.16
	20%	0.83	0.84	0.97	0.35
	30%	0.90	0.91	0.97	0.49
	40%	0.93	0.94	0.98	0.58
	50%	0.94	0.95	0.98	0.59
	60%	0.94	0.96	0.97	0.61
	70%	0.94	0.96	0.97	0.61
	80%	0.95	0.97	0.97	0.62
	90%	0.96	0.97	0.98	0.63
Baseline 2	10%	0.63	0.68	0.64	0.26
	20%	0.71	0.74	0.76	0.41
	30%	0.75	0.80	0.78	0.48
	40%	0.76	0.81	0.82	0.49
	50%	0.78	0.84	0.85	0.50
	60%	0.80	0.87	0.86	0.44
	70%	0.83	0.89	0.90	0.50
	80%	0.85	0.91	0.92	0.41
	90%	0.87	0.93	0.98	0.50

Baseline1: BERT-Base-Uncased model for sequence classification
Baseline2: A simple CNN with one convolution and one fully connected layer.

Second Evaluation Scheme

- **Aim:** Test the proposed framework against the performance of the Baseline Supervised models (BERT, CNN)
- **Method:**
  - Cluster the tweets by using ktrain's zero-shot topic classification model. N different clusters
  - Train on k and test on *N*-*k*. Start 1, *N*-1; till *N*-1, 1
    - Chosen as 91 for the CoAID and 51 for the COVID-19 Rumors dataset respectively.
  - Significant class imbalance problem in favor of True posts in the CoAID dataset (7% False, 93% True posts).
  - Under-sampling on CoAID dataset applied for supervised models based on clusters
  - Early stopping is applied since the data size of training is changing

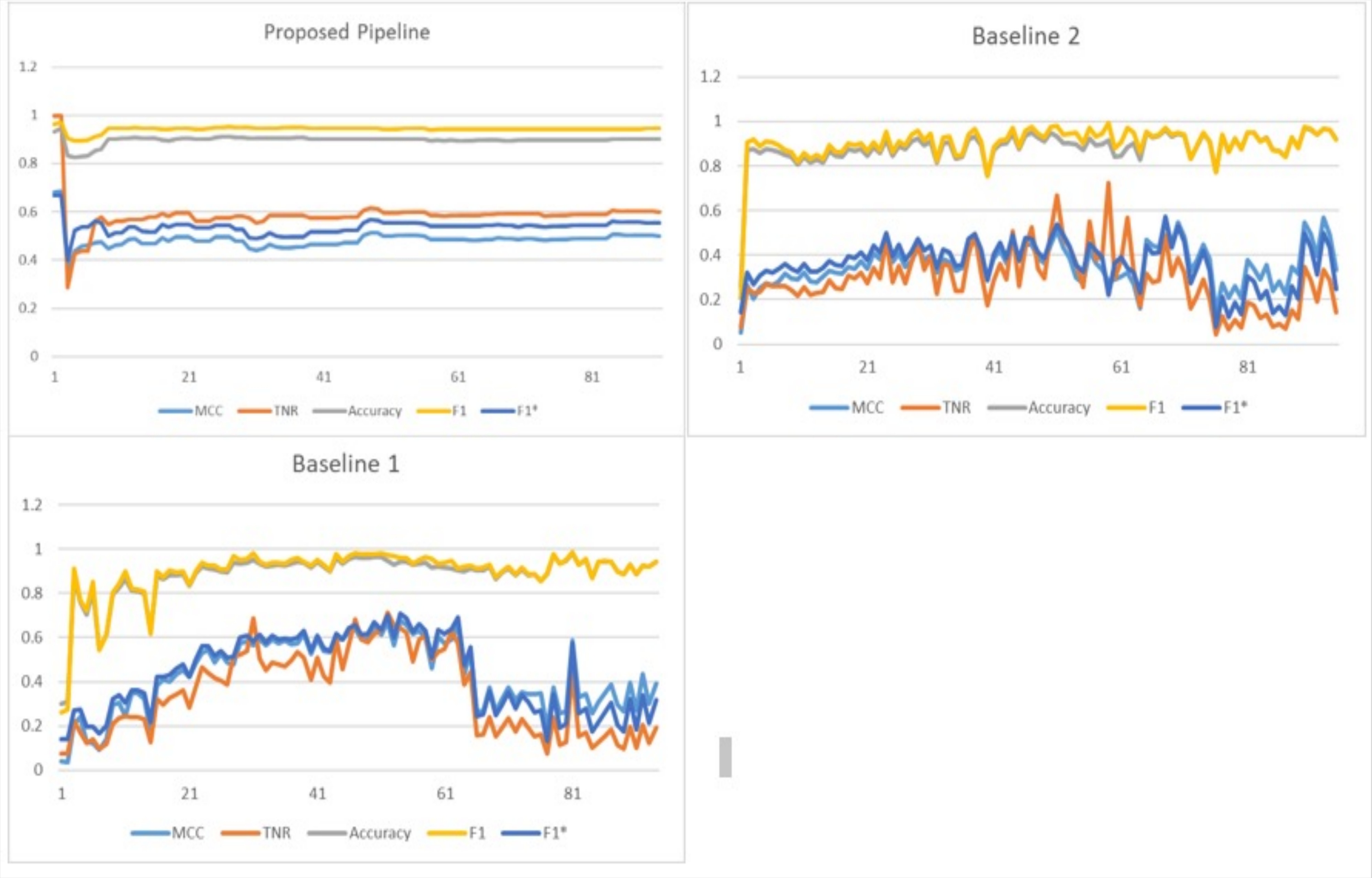


Figure 4 The testing results of the proposed pipeline and the baseline models. The vertical axis represents the value of metrics, and the horizontal axis represents the cluster count

Second Evaluation Scheme

- **Evaluation**
  - Matthews Correlation Coefficient (MCC), TNR and F1 scores are used.
- **Result:**
  - The proposed pipeline surpasses the baseline models in classifying False posts (TNR) and F1*, which gives more emphasis to predicting “False” posts, a requirement preferred for fake posts/news detection.
  - **The analysis shows that the pipeline underperforms, particularly on social media messages expressing an **opinion or including popular news** rather than **medical facts or claims**, and cannot be verified from the medical articles

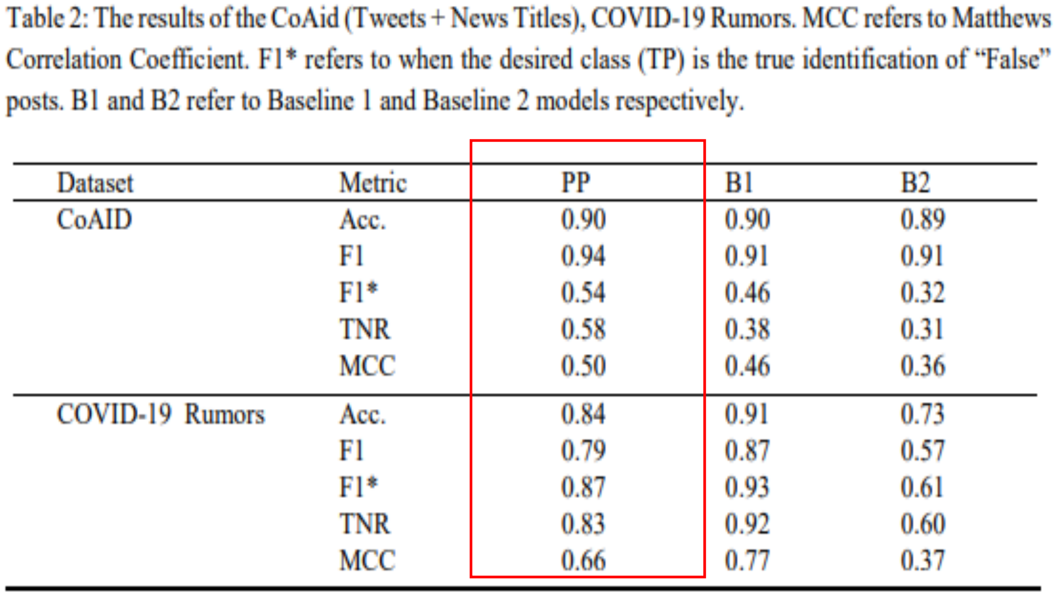


Table 2: The results of the CoAid (Tweets + News Titles), COVID-19 Rumors. MCC refers to Matthews Correlation Coefficient. F1* refers to when the desired class (TP) is the true identification of “False” posts. B1 and B2 refer to Baseline 1 and Baseline 2 models respectively.

Dataset	Metric	PP	B1	B2
CoAID	Acc.	0.90	0.90	0.89
	F1	0.94	0.91	0.91
	F1*	0.54	0.46	0.32
	TNR	0.58	0.38	0.31
	MCC	0.50	0.46	0.36
COVID-19 Rumors	Acc.	0.84	0.91	0.73
	F1	0.79	0.87	0.57
	F1*	0.87	0.93	0.61
	TNR	0.83	0.92	0.60
	MCC	0.66	0.77	0.37

Second Evaluation Scheme

- **Problematic Cases:**
  - Tweet: “*I told you guys that someone or perhaps many would die from listening to Trump and Trump’s admin. Health officials warn against self-medicating with chloroquine for coronavirus after man dies from taking fish tank cleaner*”.
  - Such opinion or daily news-related posts cannot be validated using the proposed pipeline since the pipeline checks the statements against the medical articles.
  - We conducted an experiment on user posts labeled as “Claim” only, considering those posts comprise significantly more non-medically verifiable statements than the user posts labeled as “News”. i.e., news title: “*Antiviral used to treat cat coronavirus also works against SARS-CoV-2*”, claim: “*I spent several minutes this morning chatting with the first volunteer in the Oxford COVID-19 vaccine trial via Skype*."

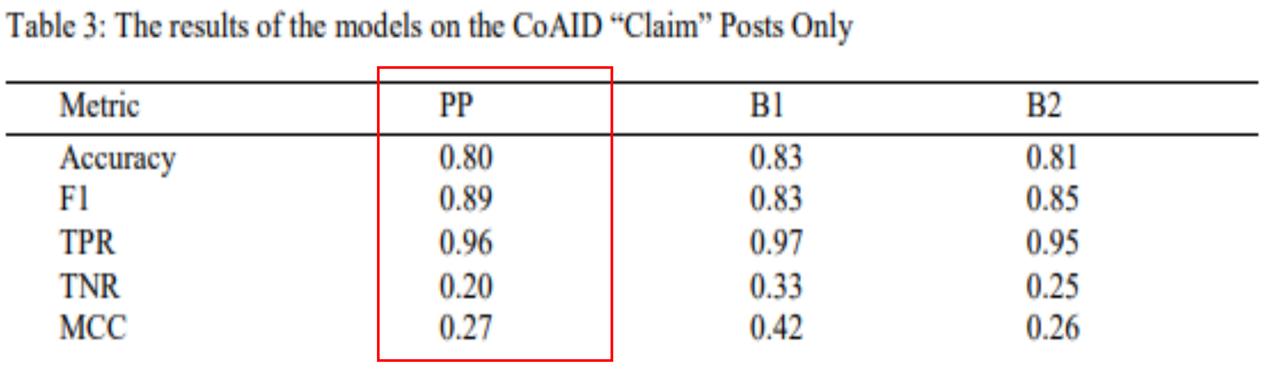


Table 3: The results of the models on the CoAID “Claim” Posts Only

Metric	PP	B1	B2
Accuracy	0.80	0.83	0.81
F1	0.89	0.83	0.85
TPR	0.96	0.97	0.95
TNR	0.20	0.33	0.25
MCC	0.27	0.42	0.26

Ablation Studies

- **Aim**:
  - To test the significance of main components in the proposed framework
  - To test research questions partially
- **Method:**
  - CoAid dataset is employed because it comprises a significant amount of medically verifiable claims
  - The ablation studies are constructed using the whole dataset
  - As the pipeline outputs three distinct labels (Supports, NotEnoughInfo, Refutes), the metrics are used for reporting multi-class classification performance specifically; accuracy, F1, Matthews Correlation Coefficient, and precision.

Ablation Studies cont.

Study 1

- **Aim:** to measure the impact of the Natural Language Inference Model chosen for the proposed pipeline.
- XLNET (M1) and BERT trained on bio-medical PubMed corpus (M2).
- **Result:**
    - BERT trained on Bio-medical data slightly outperfomed the XLNET trained on general corpora.

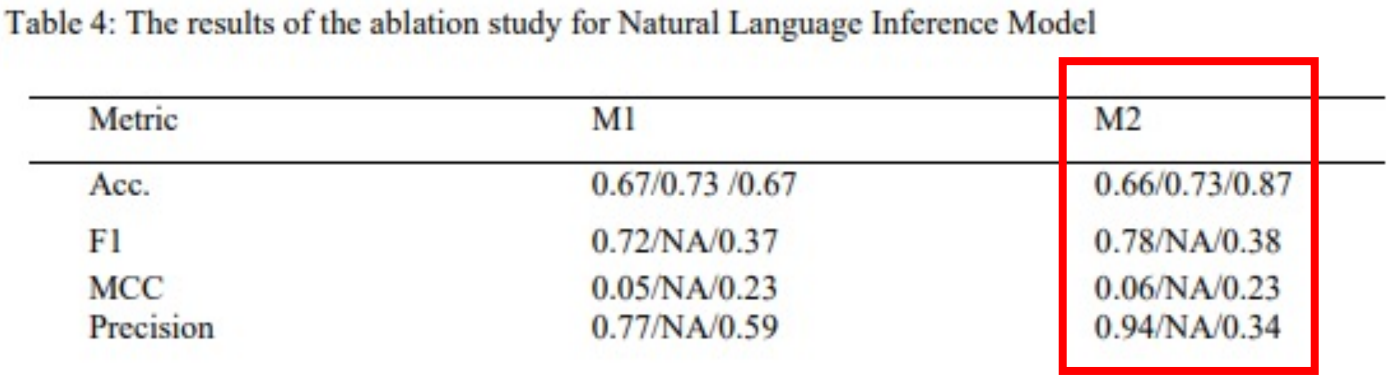


Table 4: The results of the ablation study for Natural Language Inference Model

Metric	M1	M2
Acc.	0.67/0.73 /0.67	0.66/0.73/0.87
F1	0.72/NA/0.37	0.78/NA/0.38
MCC	0.05/NA/0.23	0.06/NA/0.23
Precision	0.77/NA/0.59	0.94/NA/0.34

While the XLNET model tokenizes “*naloxone*” word as “*na-lo-xon-e*”, BERT trained on PubMed corpus tokenizes it as “*naloxone*” thus preserving the meaning of the noun and improving the results.
When the XLNET model cannot relate the words between the evidence and claim, especially in the cases where both include medical words, the pipeline tends to give “Neutral” (Not Enough Info”) as a result

Ablation Studies cont.

Study 2

- **Aim**: Can we improve medical document retrieval performance by using MeSH* (Medical Subject Headings) tree structure?
- **Result**:
  - Including search terms with the MeSH keywords significantly increases the pipeline performance
  - Summarization of evidence and check-worthy statement increases the pipeline performance moderately.

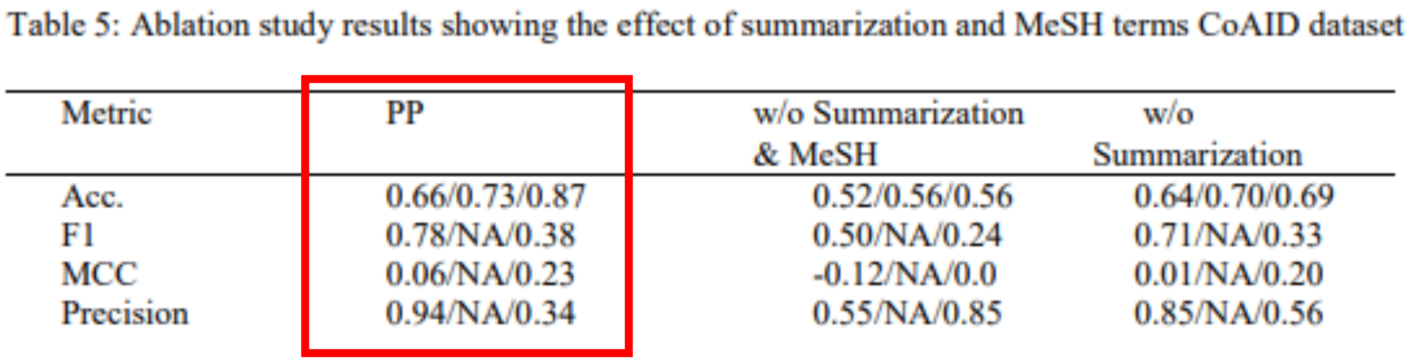


Table 5: Ablation study results showing the effect of summarization and MeSH terms CoAID dataset

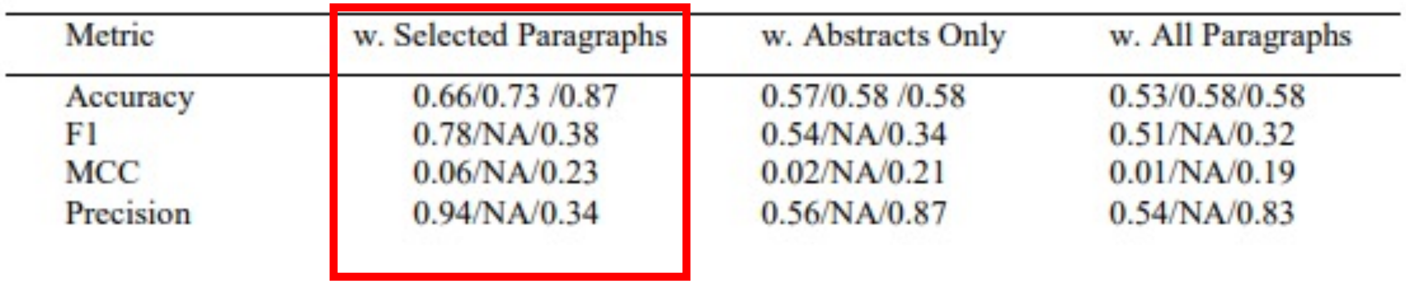
Metric	PP	w/o Summarization & MeSH	w/o Summarization
Acc.	0.66/0.73/0.87	0.52/0.56/0.56	0.64/0.70/0.69
F1	0.78/NA/0.38	0.50/NA/0.24	0.71/NA/0.33
MCC	0.06/NA/0.23	-0.12/NA/0.0	0.01/NA/0.20
Precision	0.94/NA/0.34	0.55/NA/0.85	0.85/NA/0.56

Ablation Studies cont.

Study 3

- **Aim**: Which part of the documents should we include to improve the performance? How do they affect the performance?
- **Result**:
  - Discarding sections like method, experiments and selecting only the spesified paragprahs increases the pipeline perfomance significantly.
  - Performance increment of using only the abstract, all paragraphs, and the selected paragraphs on the performance of the pipeline (Abstract, Introduction, Conclusion, Discussion and Result)

Table 6: The results of the third ablation study for evidence retrieval.



Metric	w. Selected Paragraphs	w. Abstracts Only	w. All Paragraphs
Accuracy	0.66/0.73 /0.87	0.57/0.58 /0.58	0.53/0.58/0.58
F1	0.78/NA/0.38	0.54/NA/0.34	0.51/NA/0.32
MCC	0.06/NA/0.23	0.02/NA/0.21	0.01/NA/0.19
Precision	0.94/NA/0.34	0.56/NA/0.87	0.54/NA/0.83

Discussion & Future Work

- CoAID dataset is automatically annotated, which labels tweets according to search query.
    - A better more representative dataset is needed.
- ClaimBuster was trained on the political claims. Model trained on the domain data might be better.
- Similarity score based answer/ evidence ranking algorithm**

**e.g., Claim: The new coronavirus cannot be transmitted through mosquito bites
Evidence: Sindbis virus (sinv) a positivestranded RNA virus that causes mild symptoms in humans is transmitted by mosquito bites.

- Evaluation of the performance of document retrieval and evidence selection.

Conclusion

- A new zero-shot fact extraction and verification framework for user posts related to COVID-19 against the medical articles that have the potential to be applied to other health domains.
- The system can successfully use user posts as search queries and find relevant evidences from scientific health-related articles
- Framework throws verdict (True, False, NEI) and related evidence as output.
- The framework shows comparable/ superior performance on these datasets in detecting fake information, including new emerging topics.

Conclusion

- The proposed model has the potential to be applied to different medical domains. The corpus to be searched can be replaced with any corpus using Anserini.
- There are three main hyperparameters that need to be changed for a different domain:
  1) The threshold used for claim extraction

     The default values can work reasonably well in a general corpus
  2) The confidence score used to retrieve the relevant evidence.
  3) The threshold used for heuristic verdict assignment.

  The latter two thresholds are needed as part of a question-answering system.

  We note that these parameters remain valid even in the case of posts for newly emerged topics since they are related to parameters for domain adaptation.

  In future work, the generalizability of these thresholds across different domains is planned to be investigated.