Countering False Information with Machine Learning

Branislav Pecher, Ivan Srba
Kempelen Institute of Intelligent Technologies (KInIT)

KInIT

RuleML Webinar

30th Sep 2020

WHO ARE WE?

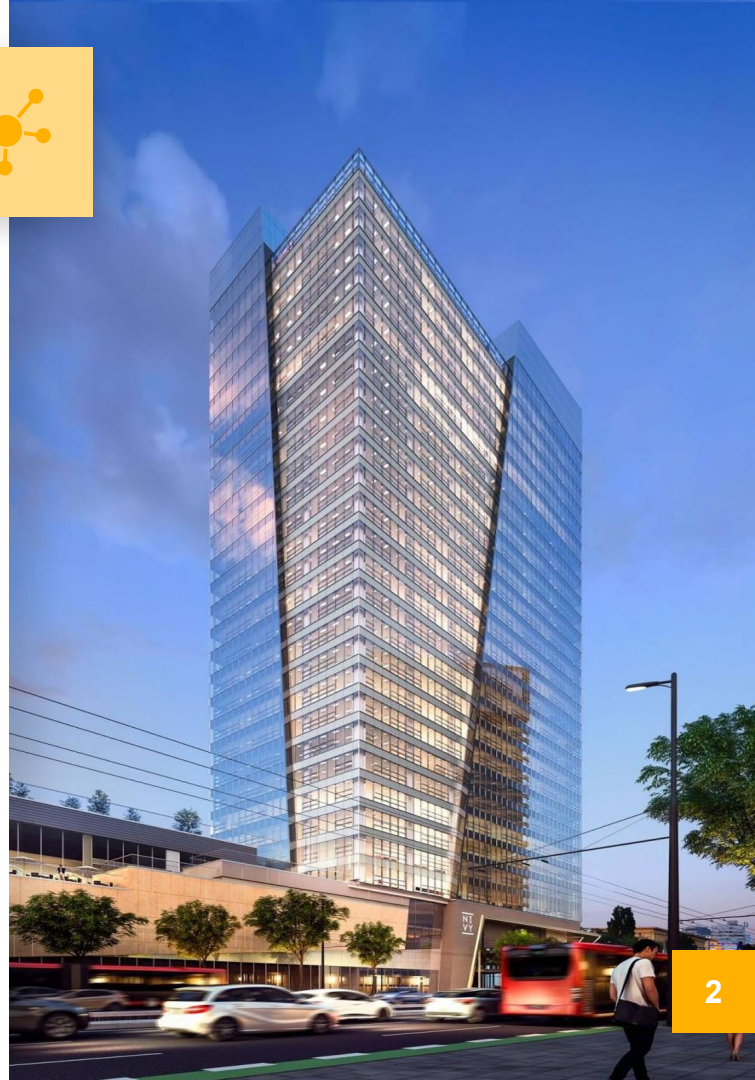
Branislav and Ivan - Researchers @KInIT

Kempelen Institute of Intelligent Technologies

- An independent institute, dedicated to research of intelligent technologies
- Located in Bratislava (Slovakia)

www.kinit.sk

KInIT



WHO ARE WE?

Kempelen Institute of Intelligent Technologies (KInIT)

Information security

Misinformation and online malicious behavior

Natural language processing

Intelligent energy grids

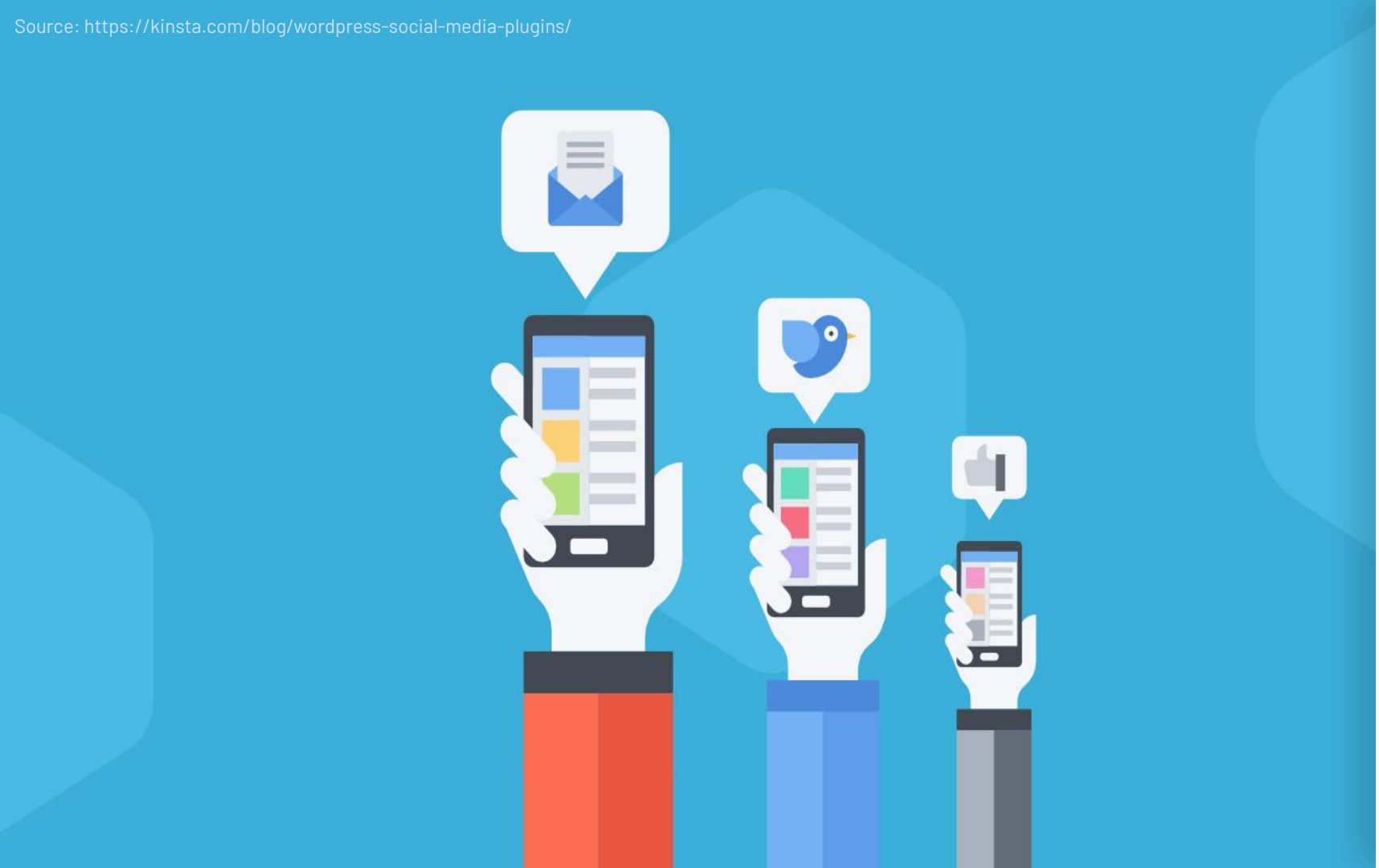
Personalization and recommendation for e-commerce

Software visualization and testing

False information

Data science perspective

Source: https://kinsta.com/blog/wordpress-social-media-plugins/



Source:
https://www.wsj.com/articles/scholars-get-the-last-laugh-on-fake-news-1515360315

Source: https://www.poynter.org/fact-checking/2019/is-expert-crowdsourcing-the-solution-to-health-misinformation/

How can

DATA SCIENCE

help to characterize, detect and mitigate false information (fake news, hoaxes, etc.)?

TASKS

Characterization

- what does characterize/distinguish, e.g., fake news from true news, how is it spread and by whom is it shared?

TASKS

Characterization

- what does characterize/distinguish, e.g., fake news from true news, how is it spread and by whom is it shared?

Detection

- how can we automatically detect fake news, etc.?

TASKS

Characterization

- what does characterize/distinguish, e.g., fake news from true news, how is it spread and by whom is it shared?

Detection

- how can we automatically detect fake news, etc.?

Mitigation

- how can we stop, e.g., the spread of fake news in a transparent, trustworthy, ethical way?

TASKS

Characterization

- what does characterize/distinguish, e.g., fake news from true news, how is it spread and by whom is it shared?

Detection

- how can we automatically detect fake news, etc.?

Mitigation

- how can we stop, e.g., the spread of fake news in a transparent, trustworthy, ethical way?

TECHNIQUES

Data analysis

Machine learning

Natural language processing

Neural networks and deep learning

Data mining

Kai Shu et al. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Expl. Newsletter 19, 1 (sep 2017), 22–36.

Srijan Kumar et al. False Information on Web and Social Media: A Survey. In Social Media Analytics: Advances and Applications. CRC press.

CHALLENGES

Labelling data (manual or automatic) is very difficult

- In some cases, it is difficult to achieve agreement also by experts
- The context is very important, a small change in wording can significantly change the meaning of the text

CHALLENGES

Labelling data (manual or automatic) is very difficult

- In some cases, it is difficult to achieve agreement also by experts
- The context is very important, a small change in wording can significantly change the meaning of the text

No suitable content-rich and benchmark datasets

https://github.com/pmacinec/fake-news-datasets

CHALLENGES

Labelling data (manual or automatic) is very difficult

- In some cases, it is difficult to achieve agreement also by experts
- The context is very important, a small change in wording can significantly change the meaning of the text

No suitable content-rich and benchmark datasets

No suitable applications and platforms to deploy solutions

https://github.com/pmacinec/fake-news-datasets

CHALLENGES

Labelling data (manual or automatic) is very difficult

- In some cases, it is difficult to achieve agreement also by experts
- The context is very important, a small change in wording can significantly change the meaning of the text

No suitable content-rich and benchmark datasets

No suitable applications and platforms to deploy solutions

Unstable terminology (e.g. fake news)

https://github.com/pmacinec/fake-news-datasets

OPEN PROBLEMS

Existing solutions focus mostly on simple (shallow) content characteristics, such as text length, text style, ...

- Vulnerable to concept drift and adversarial attacks
- Insufficient explainability

Involvement of experts or expertly prepared fact-checks

OPEN PROBLEMS

Existing solutions focus mostly on simple (shallow) content characteristics, such as text length, text style, ...

- Vulnerable to concept drift and adversarial attacks
- Insufficient explainability

  Involvement of experts or expertly prepared fact-checks

Most works focus one one system, one content modality and one language

- Contextual information is crucial for false information detection

  Multisource, multimodal and multilingual approaches

OPEN PROBLEMS

Standard supervised approaches do not explicitly address unlabelled and dynamic data

- Potential of large amount of unlabeled data remains untapped

  Semi-supervised learning, active learning, transfer learning, meta-learning

OPEN PROBLEMS

Standard supervised approaches do not explicitly address unlabelled and dynamic data

- Potential of large amount of unlabeled data remains untapped

  Semi-supervised learning, active learning, transfer learning, meta-learning

Limited research on mitigation approaches

- High demand for multidisciplinary research

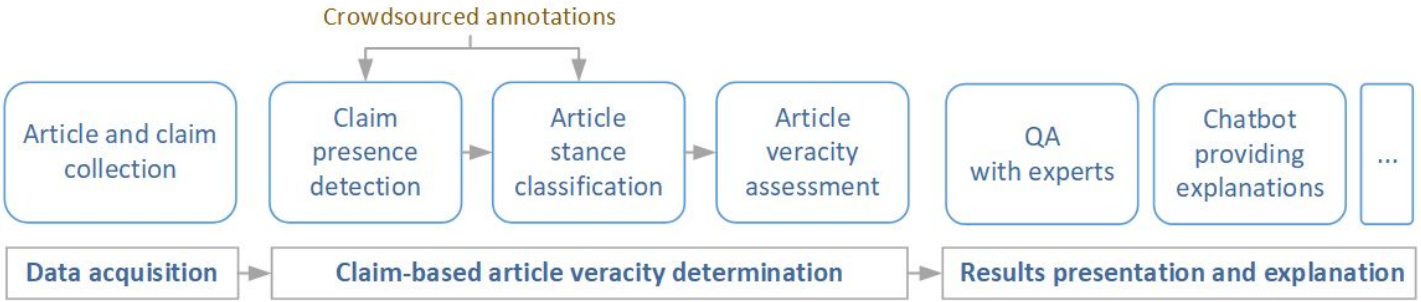
  Early warning system, on-site warning system, education and training

Misinformation detection

Based on machine learning and fact-checked claims

MISINFORMATION DETECTION

Actual content veracity determined by claims fact-checked by experts instead of simple content characteristics



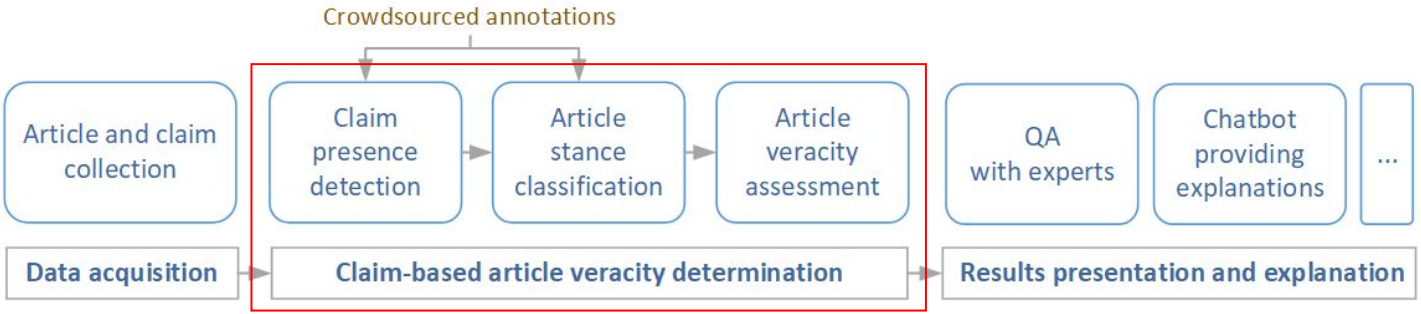
Wang et al.: Relevant Document Discovery for Fact-Checking Articles. WWW. 2018.

Popat et al.: DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. ACL. 2018.

Pecher et al.: FireAnt: Claim-based Medical Misinformation Detection and Monitoring. Demo @ ECML PKDD 2020.

MISINFORMATION DETECTION

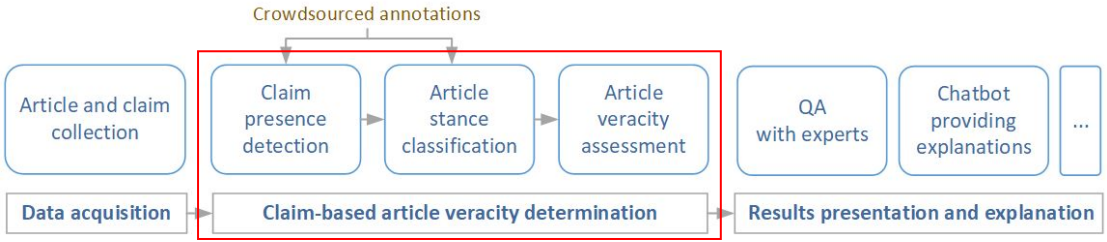
Actual content veracity determined by claims fact-checked by experts instead of simple content characteristics



Wang et al.: Relevant Document Discovery for Fact-Checking Articles. WWW. 2018.

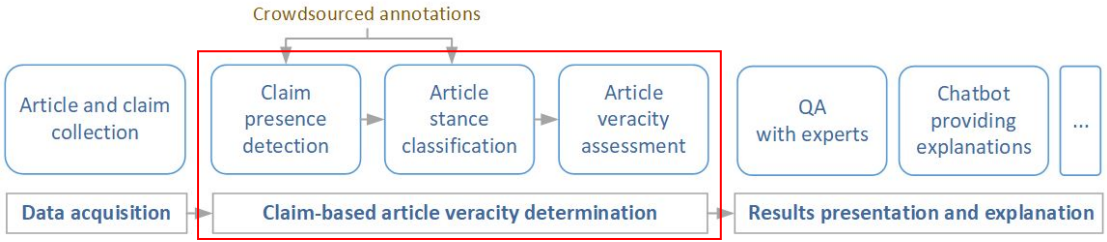
Popat et al.: DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. ACL. 2018.

Pecher et al.: FireAnt: Claim-based Medical Misinformation Detection and Monitoring. Demo @ ECML PKDD 2020.



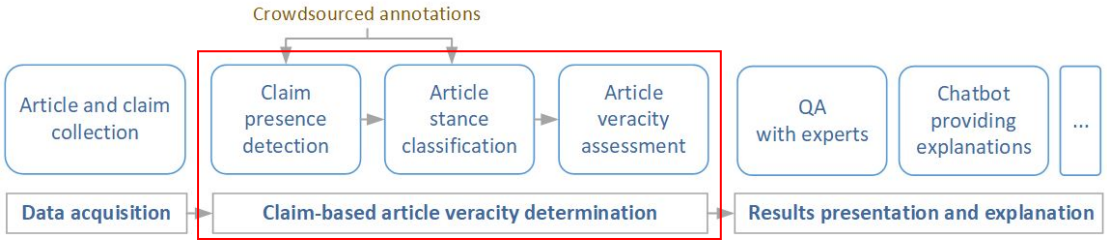
Data preprocessing

- Sentence embedding data representation
  - Universal Sentence Encoder



Data preprocessing

- Sentence embedding data representation
  - Universal Sentence Encoder
- Applied on
  - Articles (title + body sentences)
  - Claims (statement)



ARTICLE

Burning Cell Towers, Out of Baseless Fear They Spread the Virus

A conspiracy theory linking the spread of the coronavirus to 5G wireless technology has spurred more than 100 incidents this month, British officials said.

...

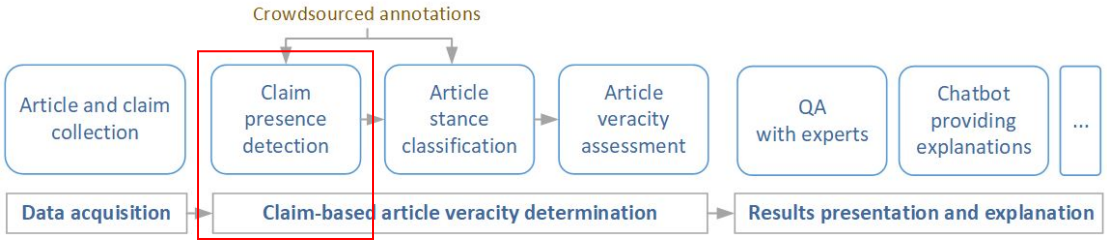
The false theory linking 5G to the coronavirus has been especially prominent, amplified by celebrities like .... on social media. It has also been stoked by a vocal anti-5G contingent, who have urged people to take action against telecom gear to protect themselves.

...

CLAIM

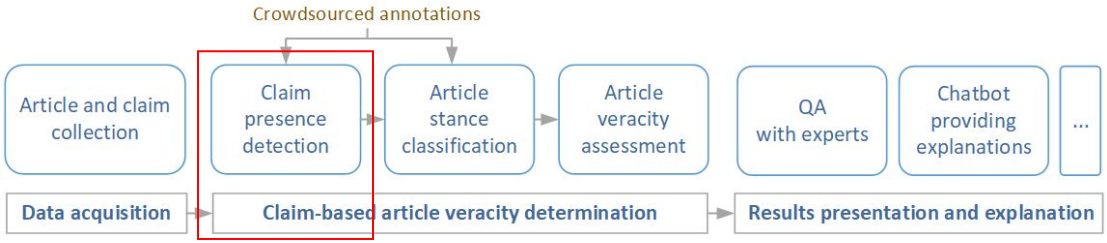
5G spreads coronavirus

Claim veracity: False



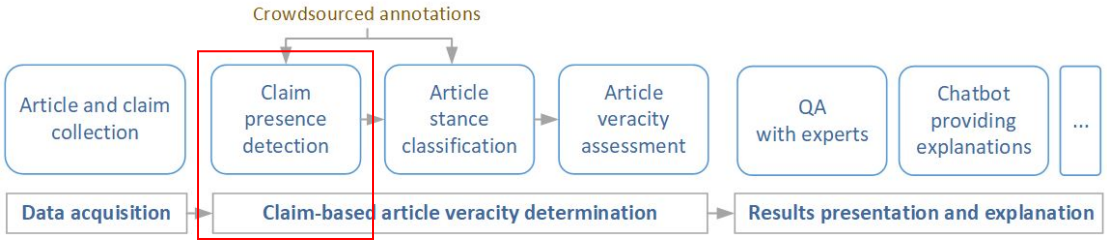
Claim presence detection

- Unsupervised IR approach
  - Into 2 classes: *present*, *not present*
- Pre-filtering step
- Matching step



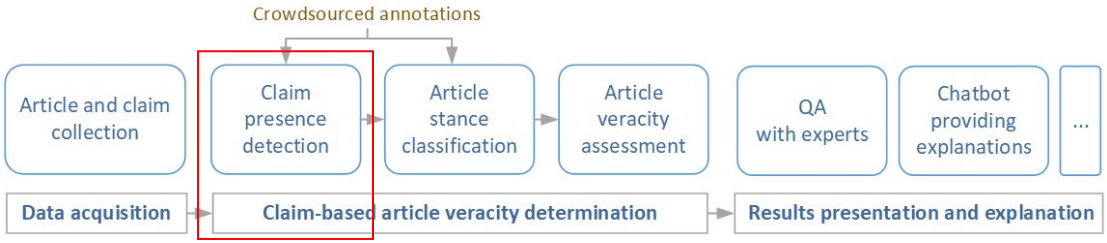
Claim presence detection

- Unsupervised IR approach
  - Into 2 classes: *present*, *not present*
- Pre-filtering step
  - Reduce number of candidates
  - Cosine similarity of title and K most similar sentences to claim
  - Compared to threshold
- Matching step



Claim presence detection

- Unsupervised IR approach
  - Into 2 classes: *present*, *not present*
- Pre-filtering step
- Matching step
  - Extract 1-, 2-, 3-grams from claim statement
  - Similarity claim-most similar sentence containing n-gram
  - n-gram TF-IDF
  - Multiplication of TF-IDF and similarity compared to threshold



ARTICLE

Burning Cell Towers, Out of Baseless Fear They Spread the Virus

A conspiracy theory linking the spread of the coronavirus to 5G wireless technology has spurred more than 100 incidents this month, British officials said.

...

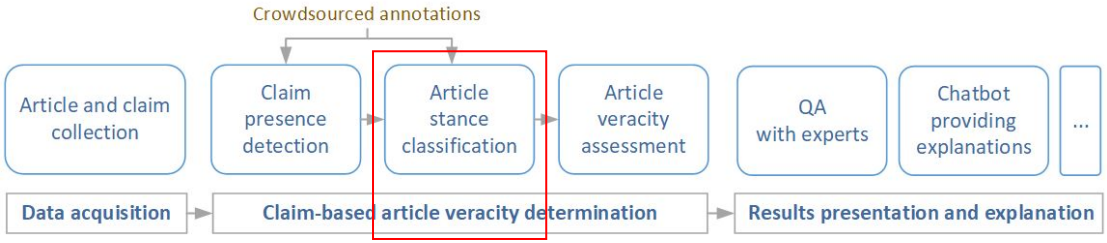
The false theory linking 5G to the coronavirus has been especially prominent, amplified by celebrities like .... on social media. It has also been stoked by a vocal anti-5G contingent, who have urged people to take action against telecom gear to protect themselves.

...

CLAIM

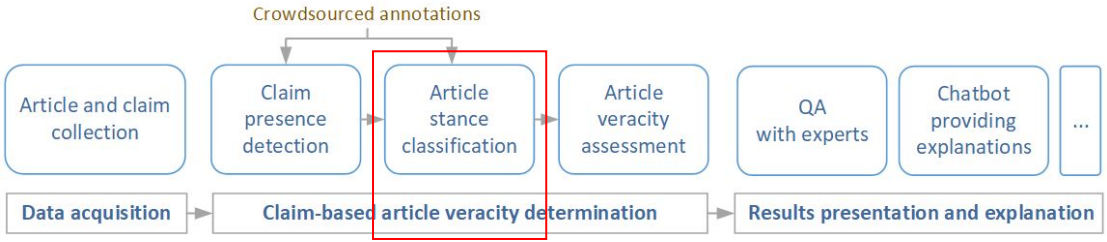
5G spreads coronavirus

Claim veracity: False



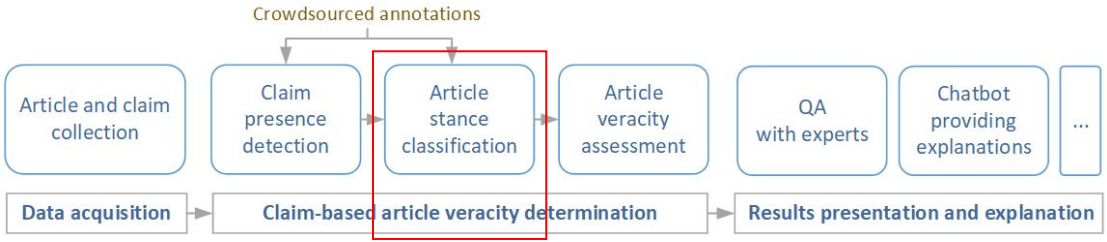
Article stance classification

- Supervised classification
  - Into 3 classes: *agree*, *disagree*, *discuss*



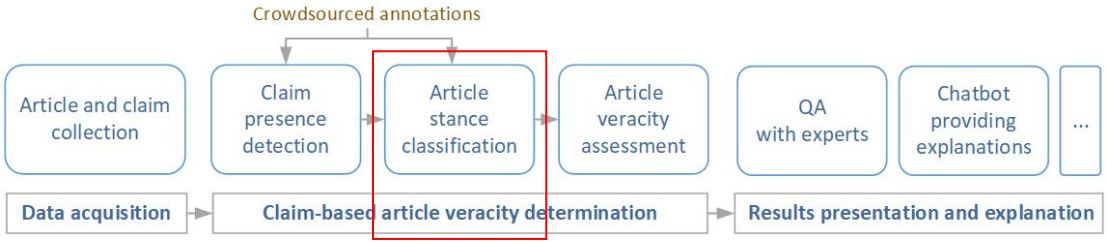
Article stance classification

- Supervised classification
  - Into 3 classes: *agree*, *disagree*, *discuss*
- Similarity CNN
  - Claim statement
  - 3 most similar sentences + their surrounding



Article stance classification

- Supervised classification
  - Into 3 classes: *agree*, *disagree*, *discuss*
- Similarity CNN
  - Claim statement
  - 3 most similar sentences + their surrounding
- Trained via transfer learning
  - Using Fake News Challenge data
  - Limited annotated data - better approach? (later)



ARTICLE

Burning Cell Towers, Out of Baseless Fear They Spread the Virus

A conspiracy theory linking the spread of the coronavirus to 5G wireless technology has spurred more than 100 incidents this month, British officials said.

...

The false theory linking 5G to the coronavirus has been especially prominent, amplified by celebrities like .... on social media. It has also been stoked by a vocal anti-5G contingent, who have urged people to take action against telecom gear to protect themselves.

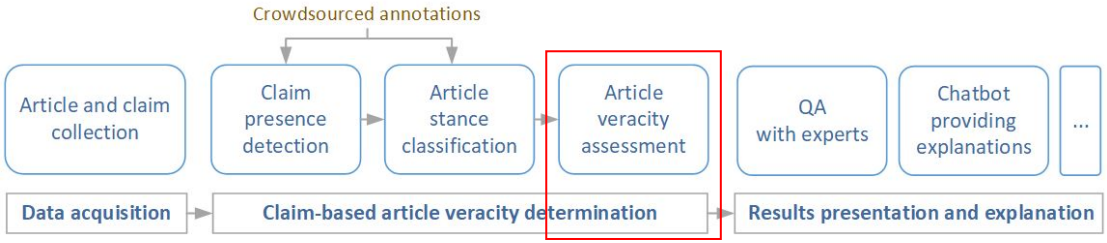
...

Article stance to the claim: Disagree

CLAIM

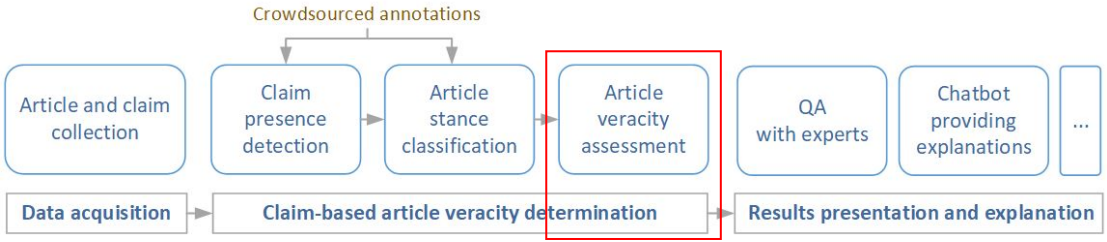
5G spreads coronavirus

Claim veracity: False



Article veracity assessment

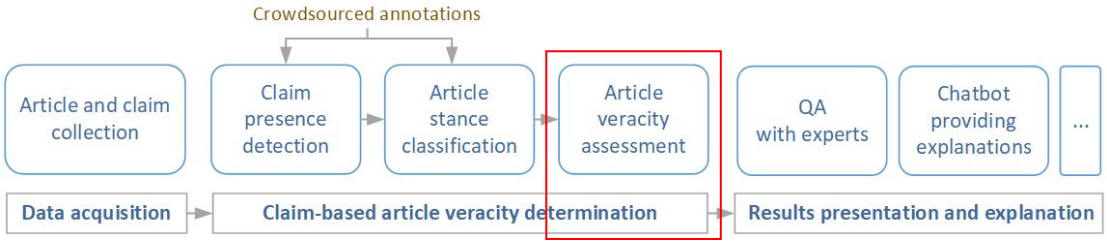
- Simple rule system
  - *false, mostly-false, mixed, mostly-true, true, unknown*



Article veracity assessment

- Simple rule system
  - *false, mostly-false, mixed, mostly-true, true, unknown*

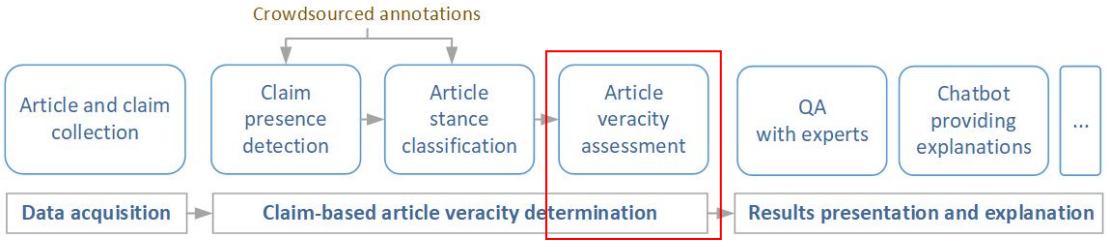
1. Article agrees with claim - Use claim veracity



Article veracity assessment

- Simple rule system
  - *false, mostly-false, mixed, mostly-true, true, unknown*

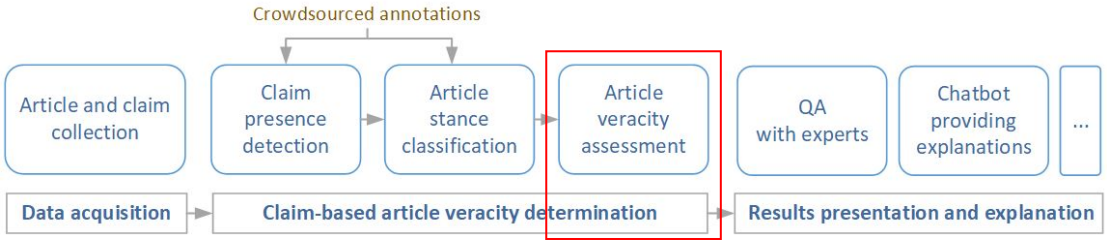
1. Article agrees with claim - Use claim veracity
2. Article disagrees with claim - Use inverse of claim veracity



Article veracity assessment

- Simple rule system
  - *false, mostly-false, mixed, mostly-true, true, unknown*

1. Article agrees with claim - Use claim veracity
2. Article disagrees with claim - Use inverse of claim veracity
3. Article discusses the claim - Use *unknown* veracity

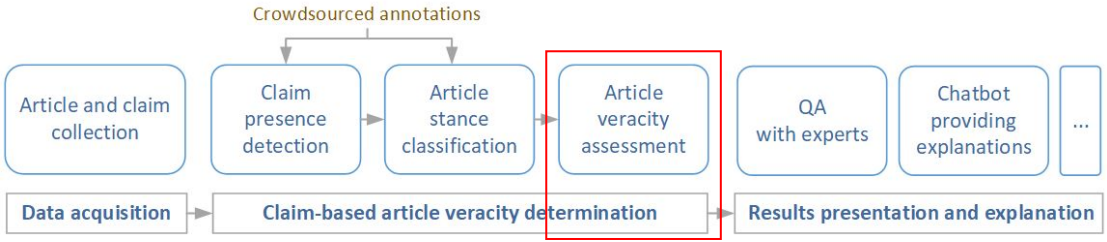


Article veracity assessment

- Simple rule system
  - *false*, *mostly-false*, *mixed*, *mostly-true*, *true*, *unknown*

1. Article agrees with claim - Use claim veracity
2. Article disagrees with claim - Use inverse of claim veracity
3. Article discusses the claim - Use *unknown* veracity

Final veracity determined as veracity of lowest rated article-claim pair



ARTICLE

Burning Cell Towers, Out of Baseless Fear They Spread the Virus

A conspiracy theory linking the spread of the coronavirus to 5G wireless technology has spurred more than 100 incidents this month, British officials said.

...

The false theory linking 5G to the coronavirus has been especially prominent, amplified by celebrities like .... on social media. It has also been stoked by a vocal anti-5G contingent, who have urged people to take action against telecom gear to protect themselves.

...

Article stance to the claim: Disagree

CLAIM

5G spreads coronavirus

Claim veracity: False

Article veracity: True

STANCE DETECTION - LIMITED ANNOTATIONS

Demanding annotation process

- Experts
- Multiple annotators to prevent mistakes

STANCE DETECTION - LIMITED ANNOTATIONS

Demanding annotation process

- Experts
- Multiple annotators to prevent mistakes

Exploit “related” data and tasks - few-shot classification

STANCE DETECTION - LIMITED ANNOTATIONS

Demanding annotation process

- Experts
- Multiple annotators to prevent mistakes

Exploit “related” data and tasks - few-shot classification

- Transfer learning

STANCE DETECTION - LIMITED ANNOTATIONS

Demanding annotation process

- Experts
- Multiple annotators to prevent mistakes

Exploit “related” data and tasks - few-shot classification

- Transfer learning
- **Meta-learning**

META-LEARNING - “LEARNING TO LEARN”

Approach with long history (1990s), repopularized recently (2016+)

META-LEARNING - "LEARNING TO LEARN"

Approach with long history (1990s), repopularized recently (2016+)

Idea: Gather and exploit previous experience

- For quick adaptation to novel tasks
- Only small number of examples used

META-LEARNING - "LEARNING TO LEARN"

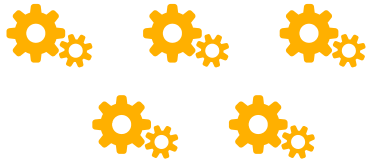
Approach with long history (1990s), repopularized recently (2016+)

Idea: Gather and exploit previous experience

- For quick adaptation to novel tasks
- Only small number of examples used

Modification of typical supervised learning

META-LEARNING - MODIFICATION OF TYPICAL LEARNING

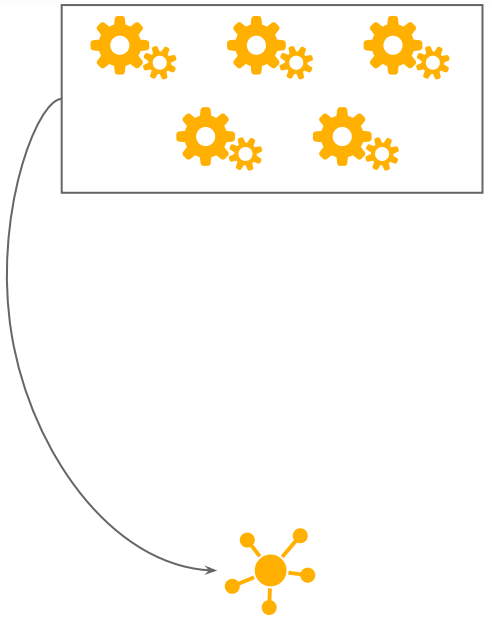


Multiple specific tasks

META-LEARNING - MODIFICATION OF TYPICAL LEARNING

Multiple specific tasks

Model gathering knowledge from task learning process

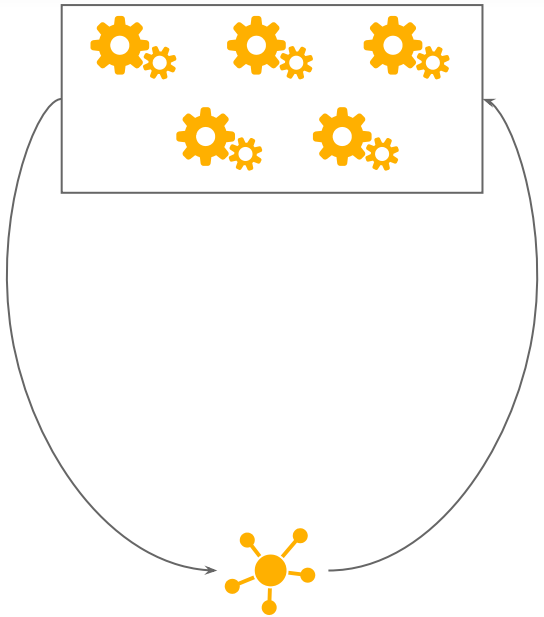


META-LEARNING - MODIFICATION OF TYPICAL LEARNING

Multiple specific tasks

Model gathering knowledge from task learning process

Improvement of specific tasks using the gathered knowledge and experience



STANCE DETECTION USING META-LEARNING

Tasks defined using different datasets

- Fake News Challenge (~20 000 samples)
- Manually annotated medical data from Monant (~190 samples)

STANCE DETECTION USING META-LEARNING

Tasks defined using different datasets

- Fake News Challenge (~20 000 samples)
- Manually annotated medical data from Monant (~190 samples)

Reptile

- SotA optimisation approach

Nichol et al.: Reptile: A Scalable Meta-Learning Algorithm. OpenAI blog. 2018.

STANCE DETECTION USING META-LEARNING

Tasks defined using different datasets

- Fake News Challenge (~20 000 samples)
- Manually annotated medical data from Monant (~190 samples)

Reptile

- SotA optimisation approach

Name / Accuracy (%)	FNC data	Monant data
Similarity CNN	65.57	56.76
Similarity CNN trained via transfer learning	71.86	74.97
Similarity CNN trained via meta-learning	**87.51**	**75.29**

Nichol et al.: Reptile: A Scalable Meta-Learning Algorithm. OpenAI blog. 2018.

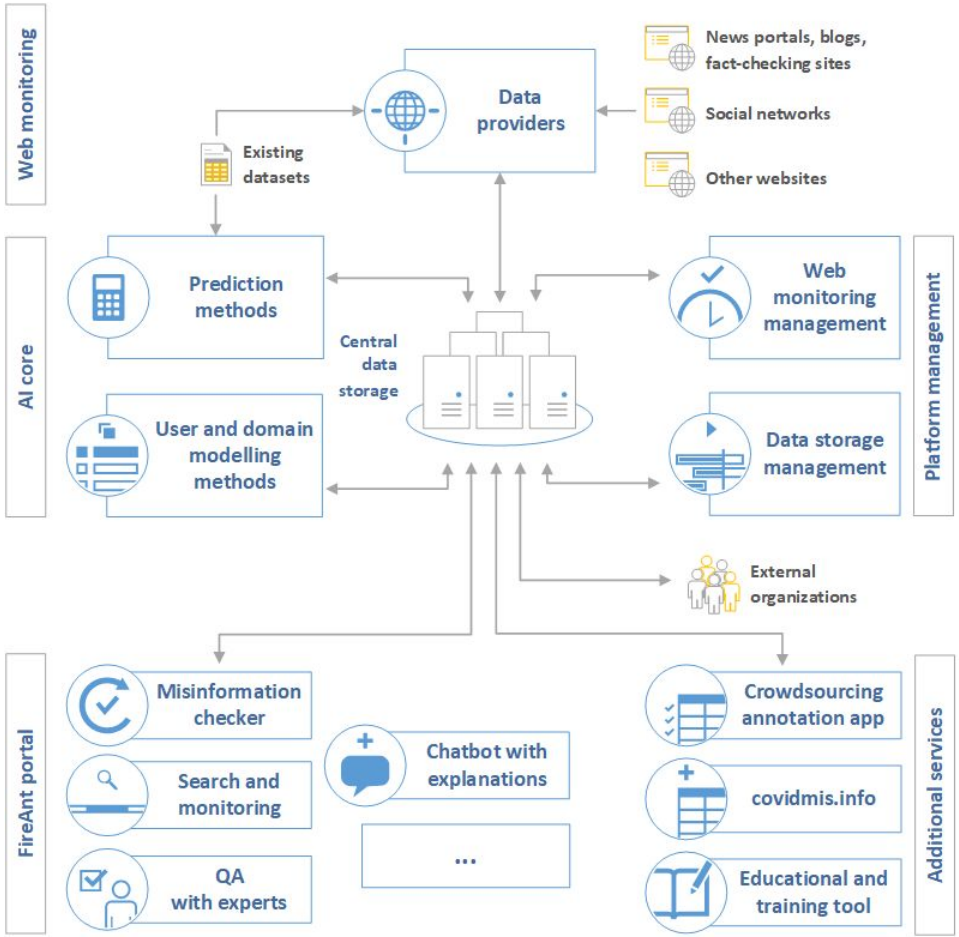
MISINFORMATION DETECTION - FUTURE WORK

Introduce meta-learning to the whole misinformation detection process

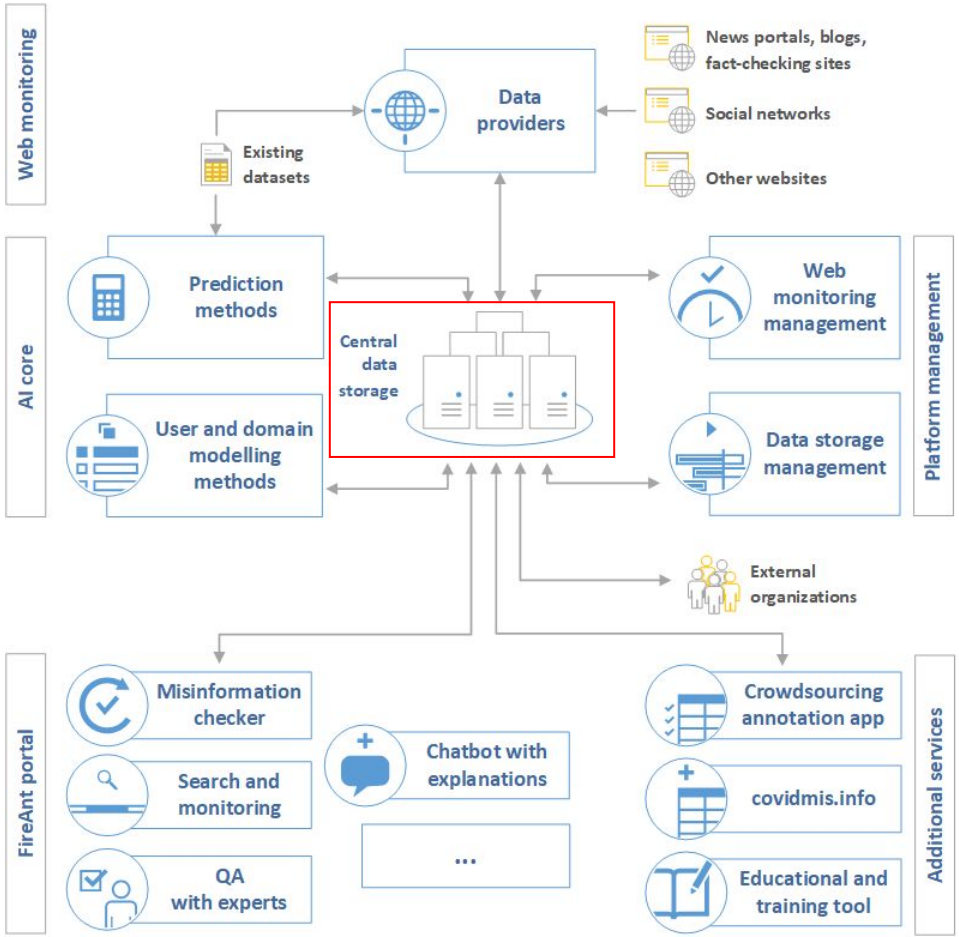
Detection across domains and languages

Monant platform

Universal and easily extensible research platform



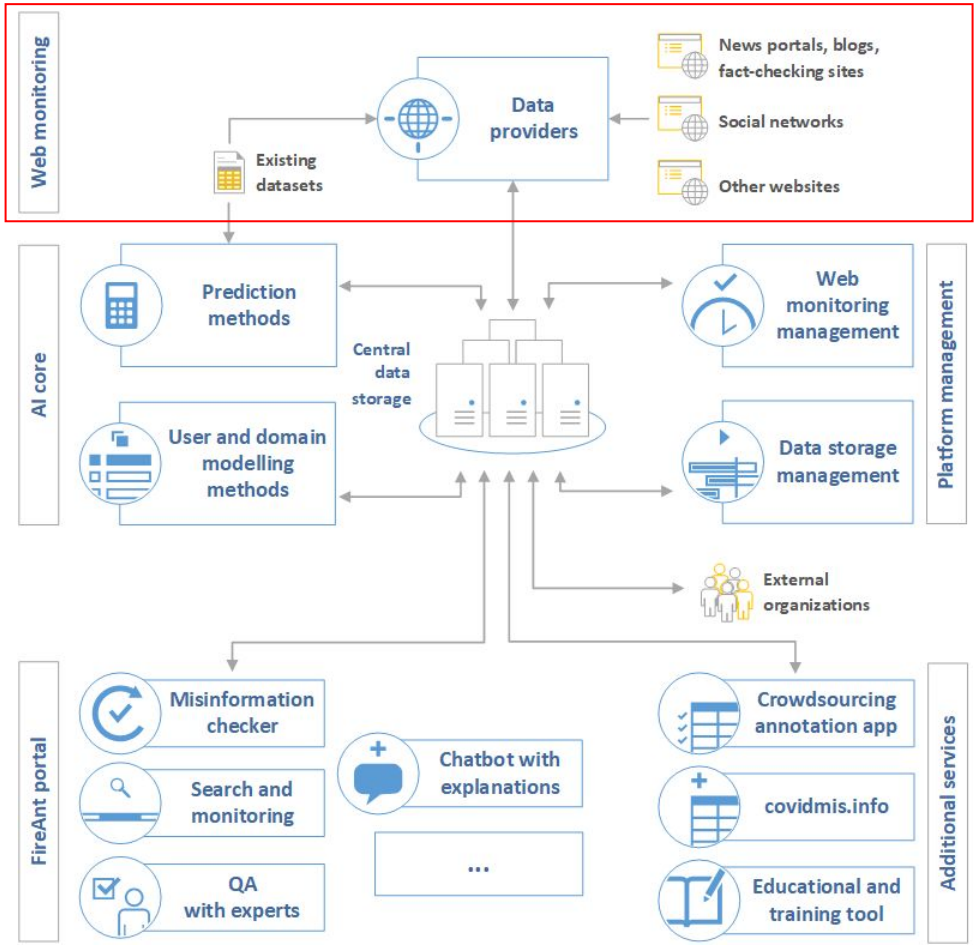
Srba et al.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. WS on Reducing Online Misinformation Exposure - ROME '19 @ SIGIR.



CENTRAL DATA STORAGE

Mediates data transfer between all platform modules by means of REST API

Srba et al.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. WS on Reducing Online Misinformation Exposure - ROME '19 @ SIGIR.

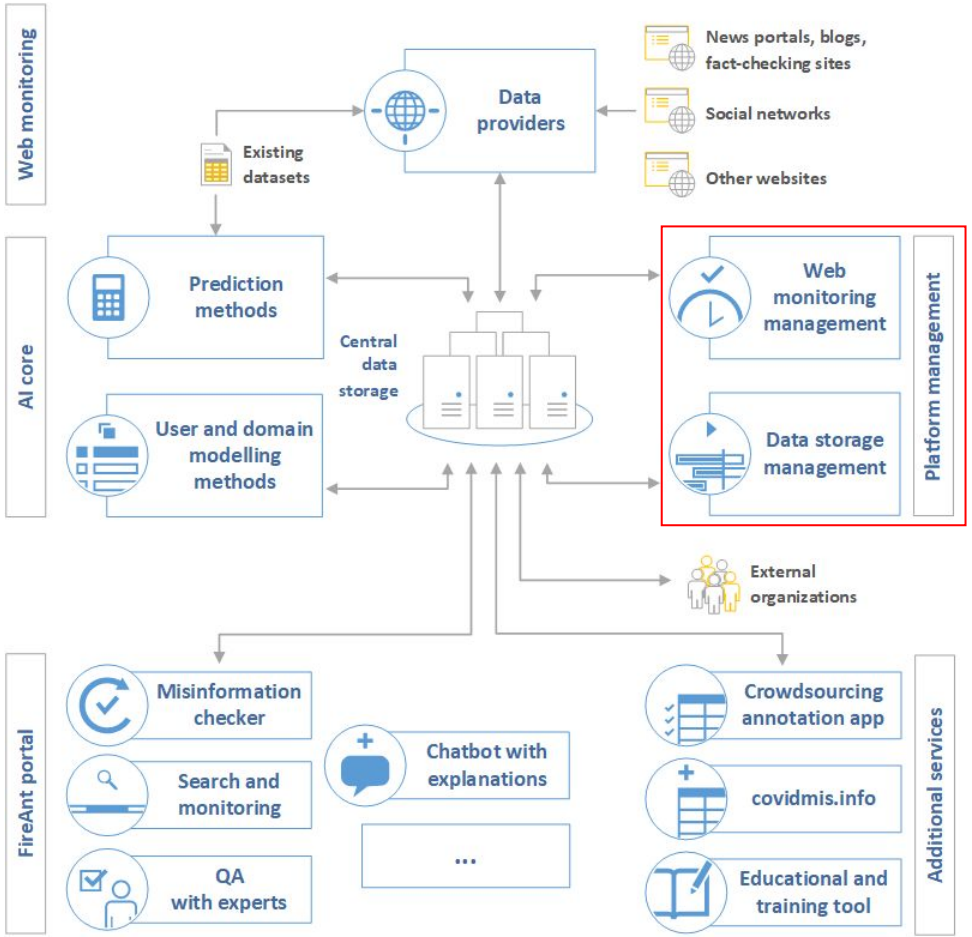


WEB MONITORING

Crawls and parses data from various data sources (news sites, fact-checking sites, etc.) by means of data providers

Event-based architecture allows chaining data providers

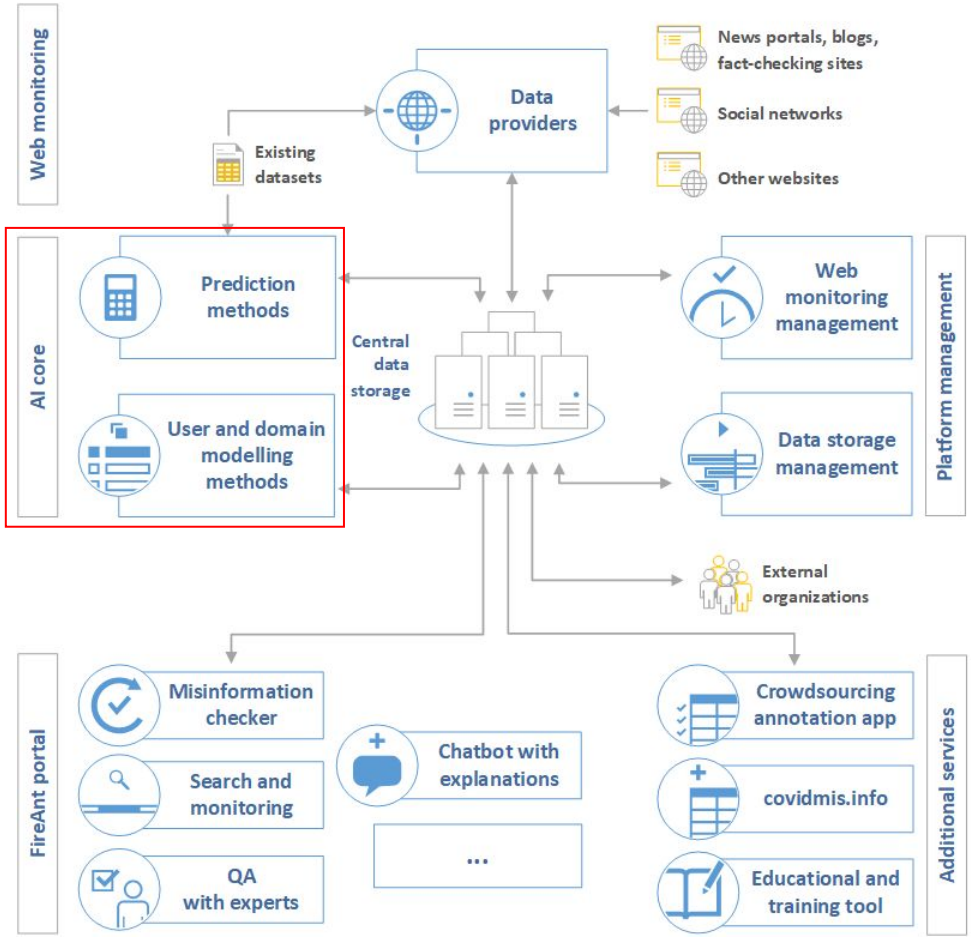
Srba et al.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. WS on Reducing Online Misinformation Exposure - ROME '19 @ SIGIR.



PLATFORM MANAGEMENT

Introduces monitors (e.g. “Monitoring health misinformation in Europe”) and access control to central data storage

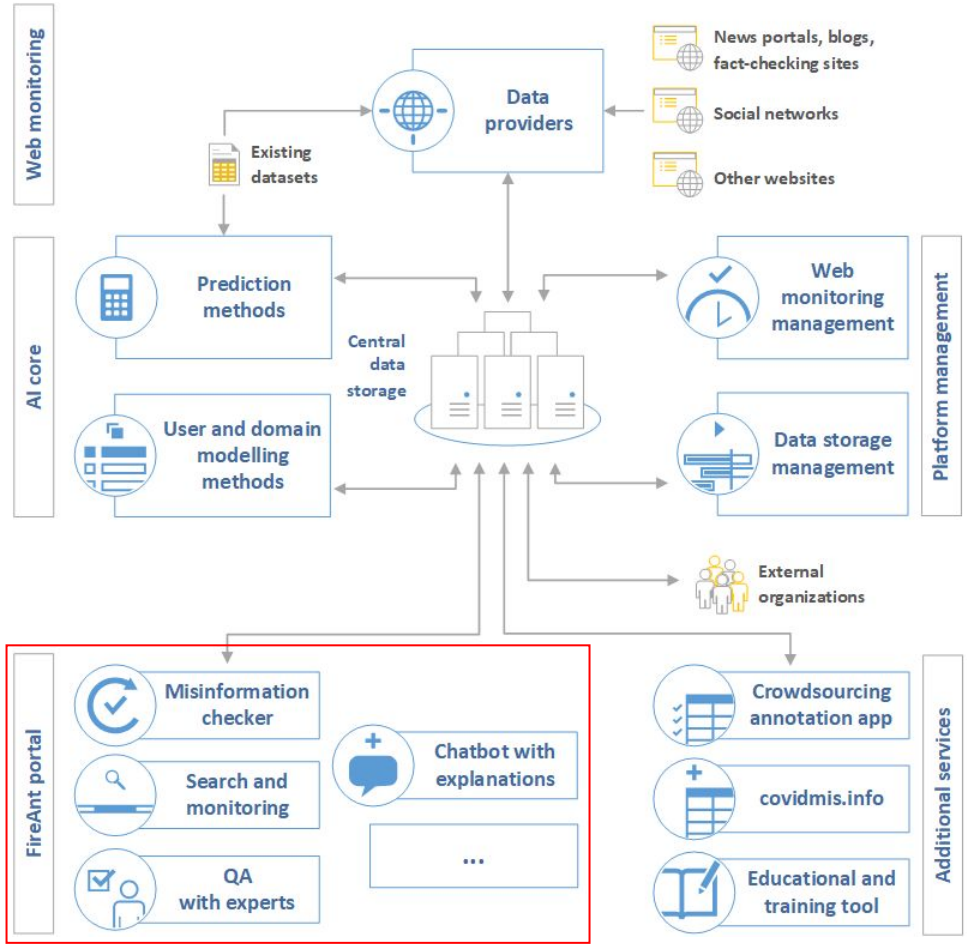
Srba et al.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. WS on Reducing Online Misinformation Exposure - ROME '19 @ SIGIR.



AI CORE

Allows to easily extend the platform with a variety of user and content modeling and prediction methods

Srba et al.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. WS on Reducing Online Misinformation Exposure - ROME '19 @ SIGIR.

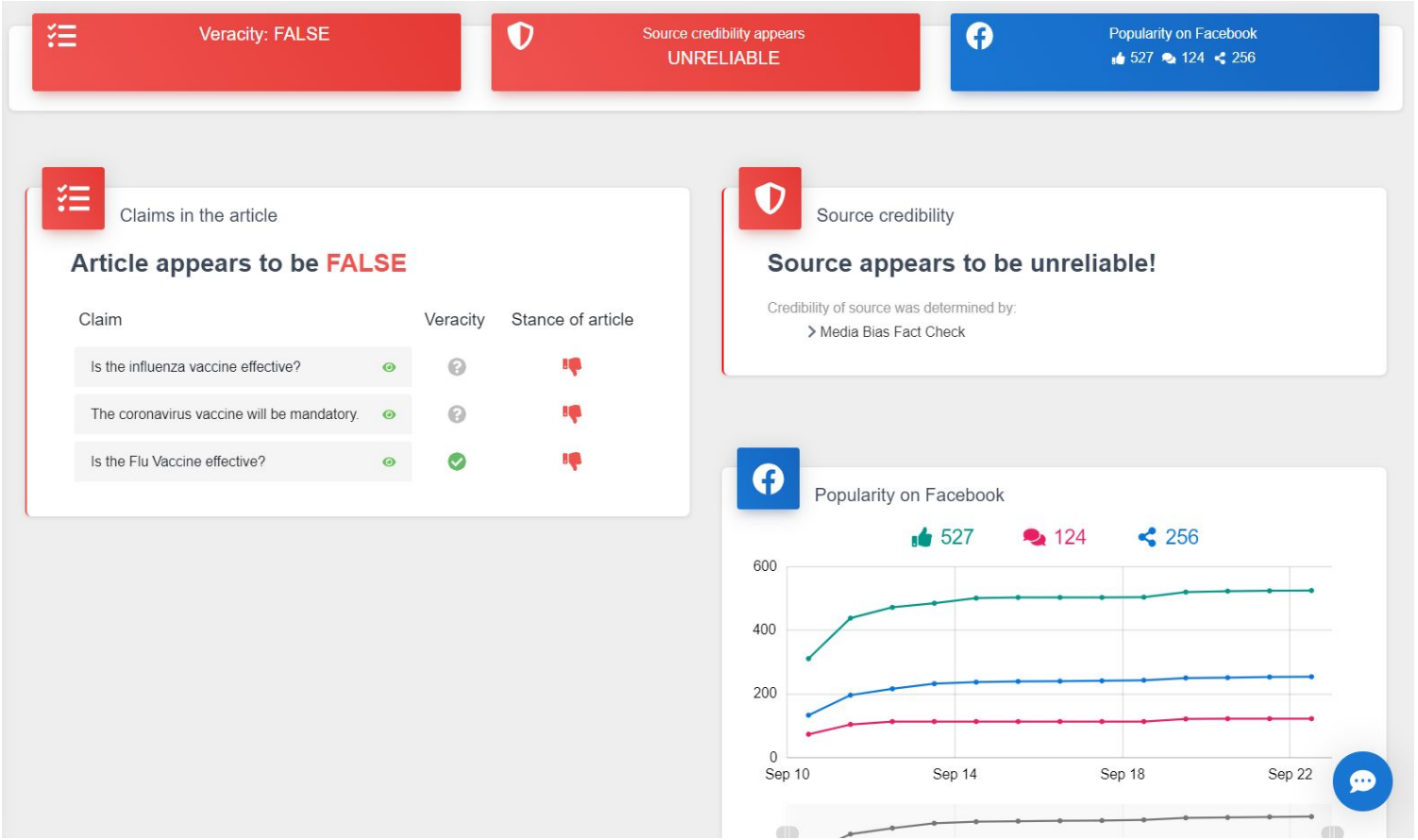


FIREANT PORTAL

Serves as an interface for experts (e.g., journalists) and general public

Provides continuously updated data, involvement of experts, explanations (e.g., by means of chatbots), ...

Srba et al.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. WS on Reducing Online Misinformation Exposure - ROME '19 @ SIGIR.



FireAnt - Article checker

MONANT DATA STATISTICS

Focus on medical articles in English, Slovak and Czech language

Sources: 200

Articles: 625 thousand
(250 thousand are English and medical, 11 thousand have a veracity label)

Claims: 8205
(2411 has a veracity already determined by experts)

Discussion posts: 624 thousands

CONCLUSIONS

False information detection have a strong negative effect on individuals and society

CONCLUSIONS

False information detection have a strong negative effect on individuals and society

Instead of shallow content features, we focus on detection based on expertly fact-checked claims

CONCLUSIONS

False information detection have a strong negative effect on individuals and society

Instead of shallow content features, we focus on detection based on expertly fact-checked claims

Advanced machine learning approaches, e.g. meta-learning

CONCLUSIONS

False information detection have a strong negative effect on individuals and society

Instead of shallow content features, we focus on detection based on expertly fact-checked claims

Advanced machine learning approaches, e.g. meta-learning

Monant addresses a lack of datasets and serves for AI method deployment making results available to end users

OUR CONTACT INFORMATION

KInIT

Branislav Pecher
(branislav.pecher@kinit.sk)

Ivan Srba
(ivan.srba@kinit.sk)

www.kinit.sk

Our selected publications

Srba et al.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. WS on Reducing Online Misinformation Exposure - ROME '19 @ SIGIR.

Pecher et al.: FireAnt: Claim-based Medical Misinformation Detection and Monitoring. Demo @ ECML PKDD 2020.

Šimko et al.: Fake News Reading on Social Media: An Eye-tracking Study. HT '19.