Unmasking COVID-19 False Information on Twitter: A Topic-Based Approach with BERT

Riccardo Cantini, Cristian Cosentino, Irene Kilanioti,
Fabrizio Marozzo, and Domenico Talia

University of Calabria

Email: rcantini@dimes.unical.it

26th International Conference on Discovery Science – DS2023, October 09-11, 2023

Outline

- Motivation and goals
- Background concepts
- Proposed methodology
- Experimental results
- Final remarks

Motivations and goals

- In today’s digital age, social media has become an integral part of our lives, revolutionizing the way we communicate, share information, and interact with the world around us.
- As the influence of social media continues to grow, **combating false information** has become a critical concern.
- False information in the age of **COVID-19**:
  - Exaggerated vaccine side effects
  - 5G microchips in vaccines
  - Lockdown-related conspiracy theories
- The primary risk is the **undermining of trust in science and authorities**, resulting in negative consequences for public health.

MAIN GOALS

- Identify false information in the COVID-related online conversation.
- Characterize false information from a topical perspective.
- Quantitatively assess its impact on specific discussion topics.

Background concepts

False information

- **Misinformation**: it represents inaccurate or false information shared without deliberate intent to deceive. Often a result of misunderstanding, errors, or lack of knowledge.
- **Disinformation**: it is the deliberate spreading of false or misleading information with the intent to deceive or manipulate. Typically used for propaganda and political manipulation.

- Latest approaches for false information detection include:
  - The use of deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs).
  - Transfer learning with Large Language Models (e.g., BERT).
  - Combination of transformer-based models with CNNs or RNNs.

Background concepts

Topic modeling

- Natural Language Processing techniques for uncovering latent topics (i.e., thematic structures) from large text corpora.

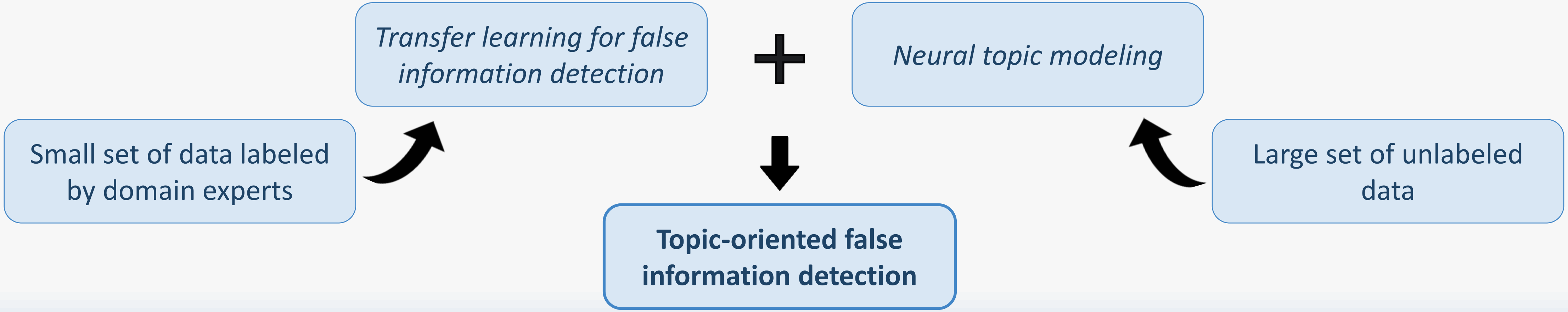
- Primary applications include document clustering and categorization, information retrieval, trend detection, and recommendation systems.

- Main techniques:
  - *Algebraic*: Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF).
  - *Probabilistic*: probabilistic-LSA (p-LSA), Latent Dirichlet Allocation (LDA).
  - *Neural-based*: LDA2Vec, Top2Vec, BERTopic.

Proposed methodology

Topic-oriented false information detection

- State-of-the-art approaches handle the false information problem within the large and comprehensive scope of COVID-19 discussions as a single entity.
- In contrast, our methodology follows a semi-supervised approach, by combining *transfer learning for false information detection* and *neural topic modeling*, to achieve a **topic-oriented representation of false information**.



Transfer learning for false information detection

+

Neural topic modeling

Small set of data labeled by domain experts

Large set of unlabeled data

Topic-oriented false information detection

Proposed methodology

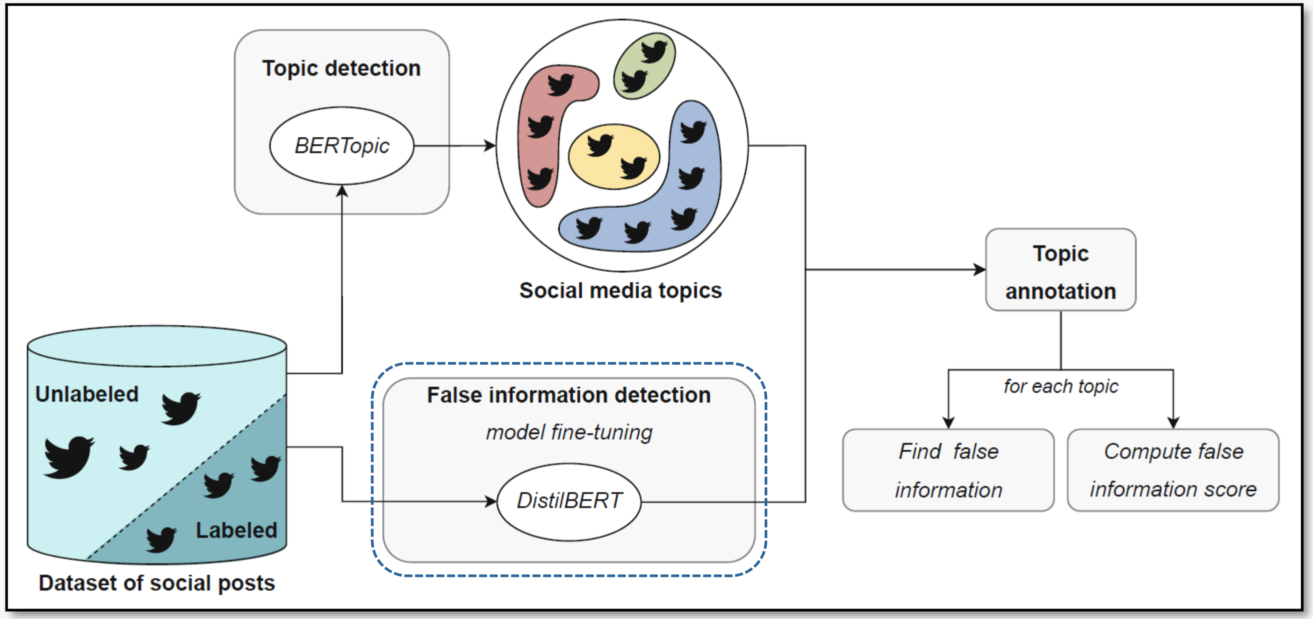
Main benefits

- By taking a topical perspective, our methodology allows for a finer-grained analysis:
  - It enables a topic-oriented quantification of the impact of false information on the online discourse surrounding COVID-19.
  - It allows for identifying the main discussion topics that are most affected by false information, also finding concrete examples of false information related to these topics.
  - It facilitates analysis of the impact of false information on specific topics, shading light on how it may have affected and shaped the online discussion at the topic level, within the broader context of COVID-19.

Proposed methodology

Fine-tuning of the false information detection model

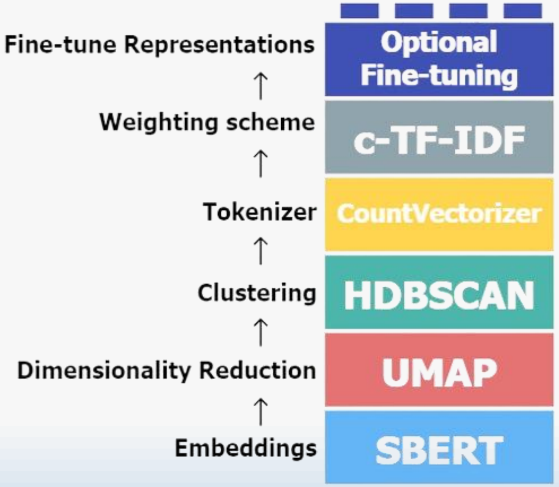
- A **DistilBERT** model is fine-tuned for the false information detection task.
- Starting from a small set of COVID-related labeled posts (~15k), a binary classifier is trained via transfer learning.
- Posts were manually annotated by health medical experts as *false information* and *reliable content*.
- In addition to evaluating DistilBERT, our testing encompassed various BERT-like models, such as BERT, ALBERT, BERTWEET, and ROBERTA.



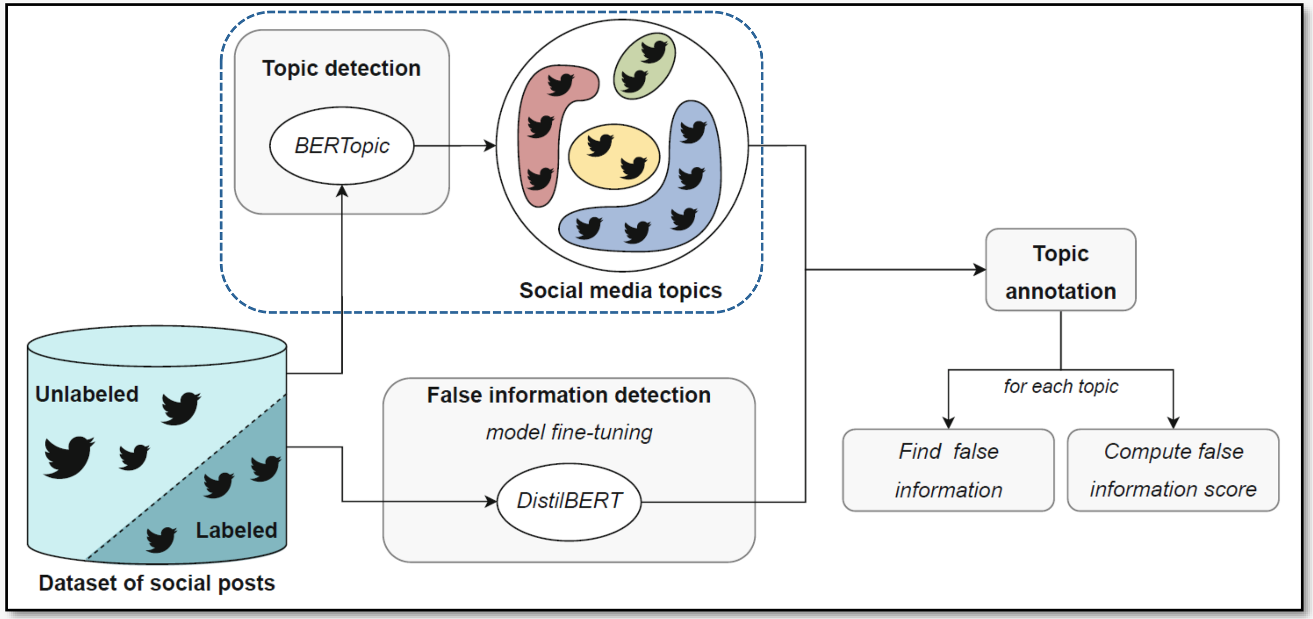
Proposed methodology

Topic detection

- **BERTopic** is leveraged to uncover the main discussion topics underlying the online discourse surrounding COVID-19.
- This step is performed on a huge set of COVID-related unlabeled posts (~300k).



*A cluster of semantically related sentences, represented in a latent space, identifies a **topic**.*



Proposed methodology

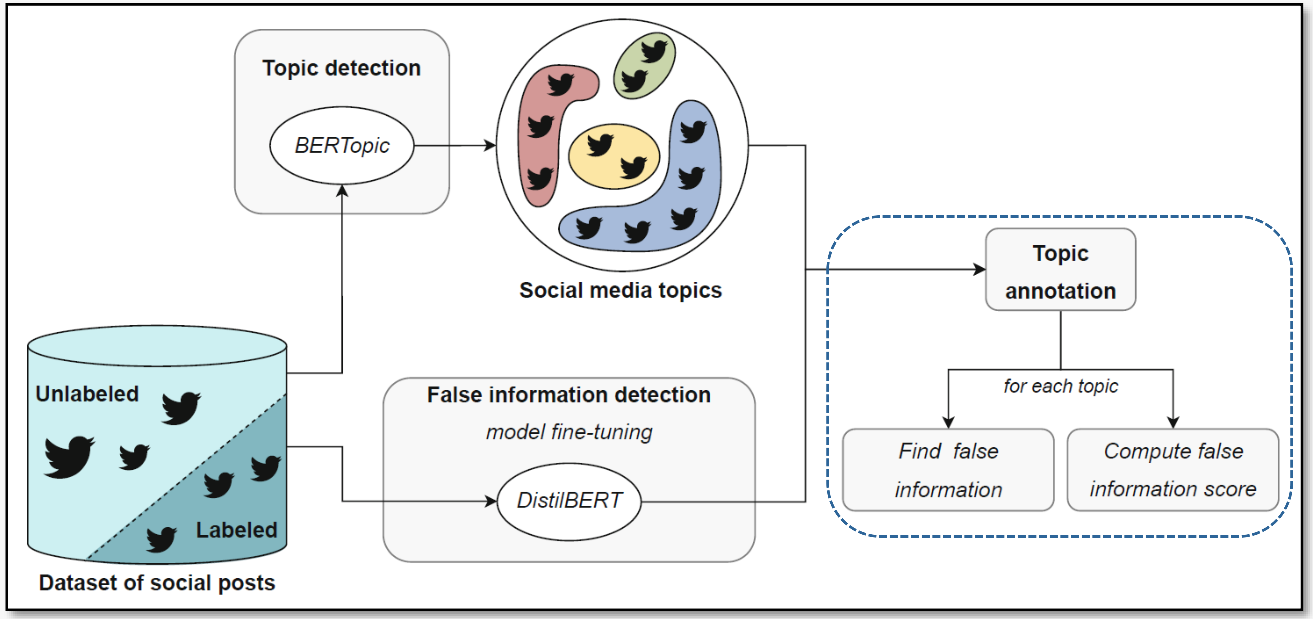
Topic annotation

- The false information detection model fine-tuned previously is used to estimate the level of false information present in each topic.

- Given a cluster *c* (i.e., a topic), a false information score is computed as follows:

$$\mathcal{S}(c) = \frac{\sum_{s \in c} p_s^c \cdot p_s^{fi}}{\sum_{s \in c} p_s^c}$$

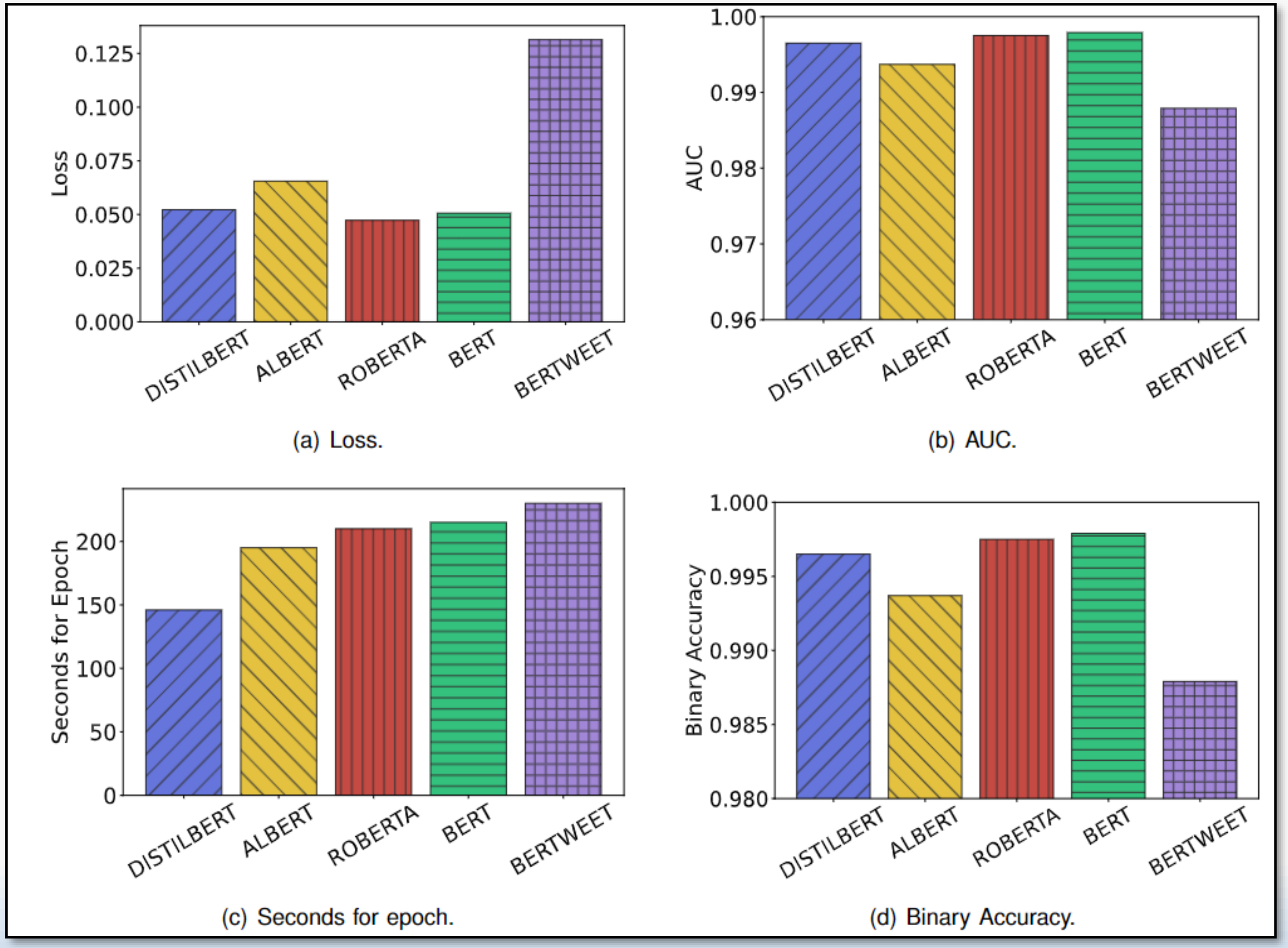
- It represents the average false information of the sentences contained in the cluster *c* (*soft labels given by **DistilBERT***), weighted on the degree of membership of those sentences (*given by **BERTopic***).



Experimental results

Model selection for false information detection

- Compared models include BERT and the following variants:
  - *DistilBERT* (knowledge distillation with ~40% fewer parameters)
  - *ALBERT* (inter-sentence coherence loss, factorized embedding parameterization)
  - *BERTWEET* (Twitter-specific variant of BERT)
  - *ROBERTA* (dynamic masking for language modeling, no NSP)
- The **DistilBERT** model achieves the best trade-off between accuracy and training time.



(a) Loss.

(b) AUC.

(c) Seconds for epoch.

(d) Binary Accuracy.

Experimental results

COVID-related detected topics

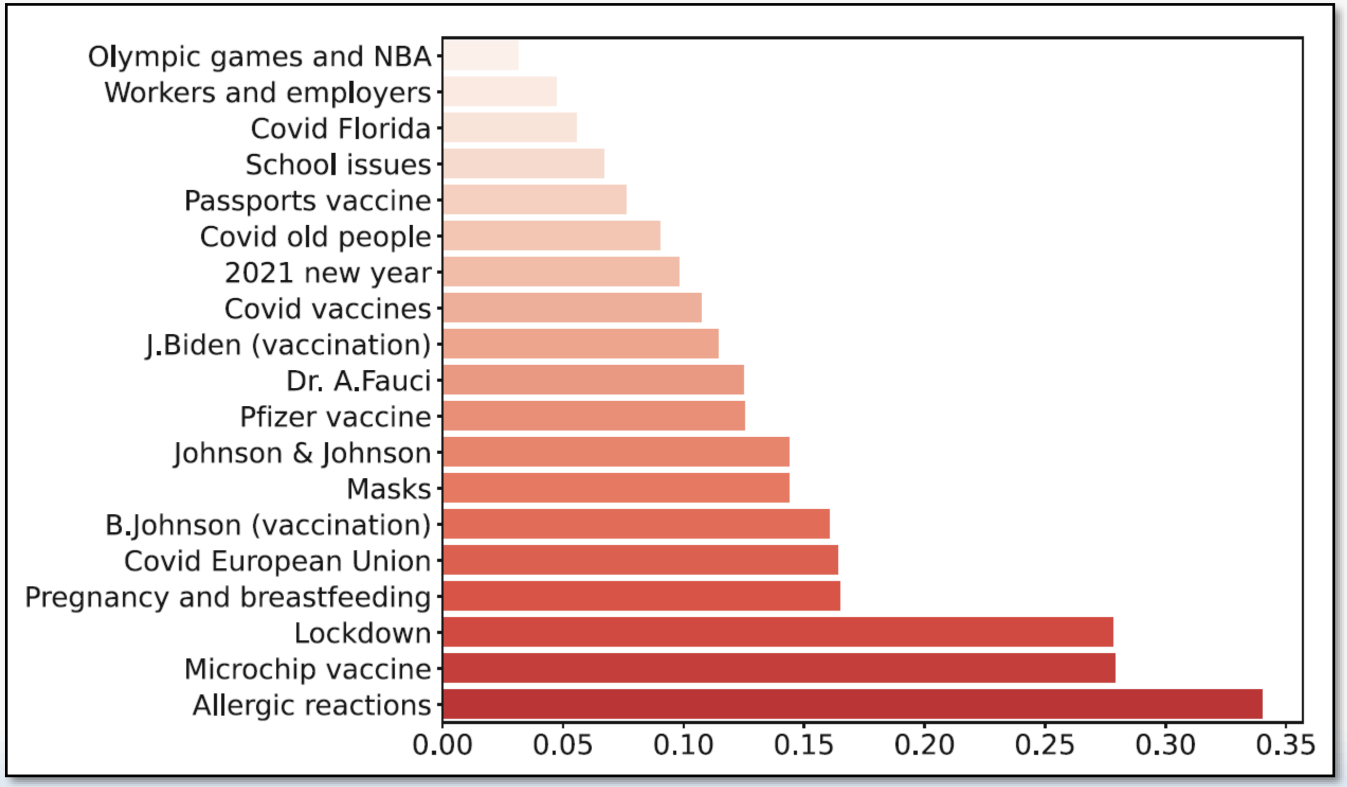
- Main identified topics include:
  - The strategy employed by the European Union and prominent politicians for handling the pandemic.
  - The effectiveness of specific vaccines (e.g., Pfizer and Johnson & Johnson).
  - Allergic reactions and risks related to pregnancy and breastfeeding.
  - The impact of COVID on workers, schools, and major sporting events like the Olympic Games.
  - Long-term effects of COVID, especially on older individuals.
  - Conspiracy theories about lockdown and 5G microchips inside vaccines.

- Topic evaluation:
  - *Coherence*: CV = 0.51, Normalized Pointwise Mutual Information = 0.09
  - *Diversity*: Percentage of Unique Words = 0.97, average pairwise Jaccard Distance = 0.99

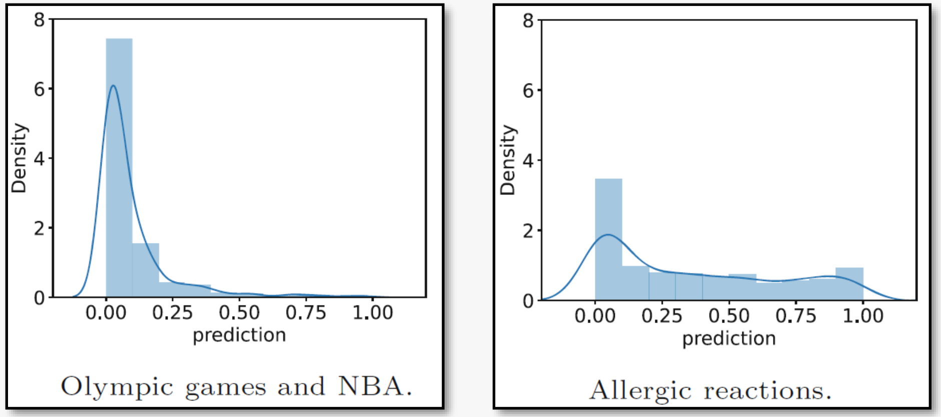
Experimental results

Topic-oriented false information detected in COVID discussions

- False information score for each identified topic:



- DistilBERT output distribution for the topics with the **lowest** and the **highest** level of false information (i.e., *Olympic games and NBA* and *Allergic reactions*, respectively).



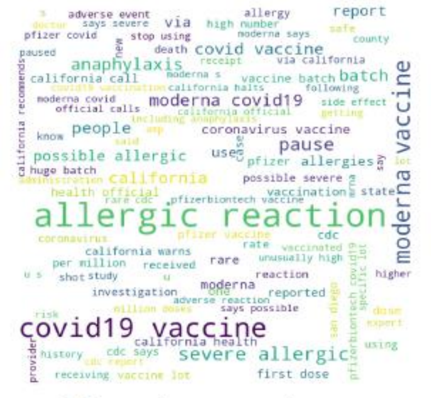
Olympic games and NBA.

Allergic reactions.

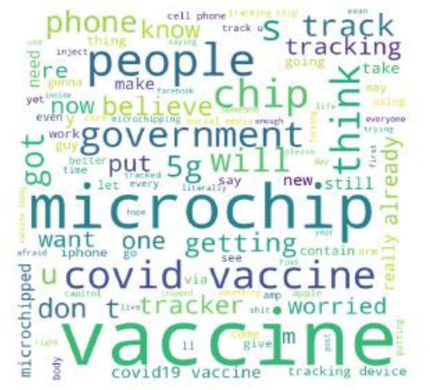
Experimental results

Top-3 topics with the highest false information score

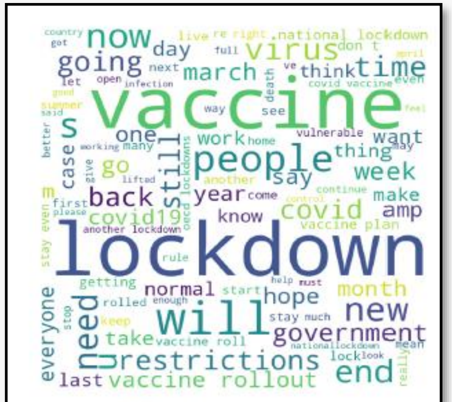
- **Allergic reactions**: this topic refers to side effects that may occur after receiving a vaccine injection, but also contains fake news about presumed and/or exaggerated vaccine side effects and no-vax instigations.
  - False information score: 0.35.
- **Microchip vaccine**: the topic refers to conspiracy theories regarding the presence of 5G microchips in vaccines.
  - False information score: 0.25.
- **Lockdown**: the topic refers to lockdown measures imposed during the COVID-19 pandemic, but also contains discussions about lockdown-related conspiracy theories.
  - False information score: 0.24.



Allergic reactions.



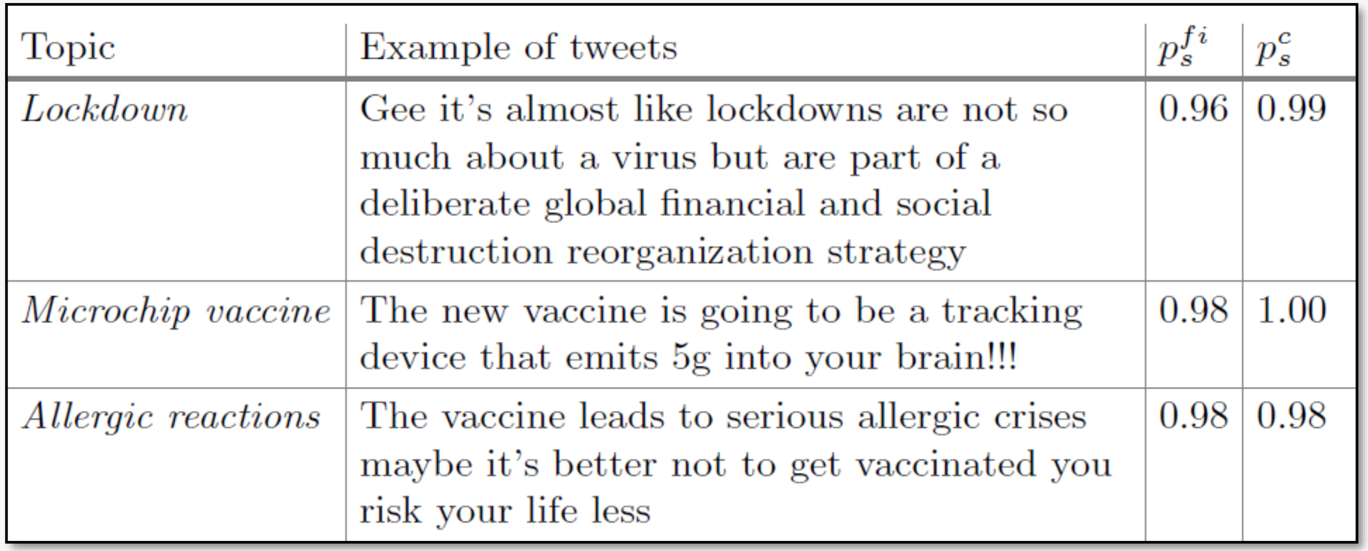
Microchip vaccine.



Lockdown.

Experimental results

Top tweets per false information



Topic	Example of tweets	p_s^{fi}	p_s^c
Lockdown	Gee it's almost like lockdowns are not so much about a virus but are part of a deliberate global financial and social destruction reorganization strategy	0.96	0.99
Microchip vaccine	The new vaccine is going to be a tracking device that emits 5g into your brain!!!	0.98	1.00
Allergic reactions	The vaccine leads to serious allergic crises maybe it's better not to get vaccinated you risk your life less	0.98	0.98

- The reported tweets express:
  - skepticism and concerns about lockdown measures and COVID-19 vaccines;
  - conspiracy theories, including the belief that vaccines emit harmful 5g waves and contain surveillance microchips;
  - unsupported severe side effects of vaccines and no-vax instigations.

Final remarks

- Our work focuses on the analysis of Twitter conversations to uncover and address false information pertaining to **COVID-19**.
- Instead of facing false information as a single entity, our work takes a topical perspective by combining *transfer learning for false information detection* and *neural topic modeling*, to achieve a **topic-oriented representation of false information**.
- Following this approach, our study provides detailed information about COVID-related false information, allowing us to quantitatively assess the presence of false information in the main topics discussed by social users.

Final remarks

- Among the topics with the highest incidence of false information, we found allergic reactions and conspiracy theories related to lockdown measures and 5G microchips in vaccines.

- A precise and finer-grained identification of the main sources of false information can enhance the reliability and trustworthiness of content shared on social media, ultimately benefiting society and public health.

- As a future direction, we will investigate the integration of *dynamic topic modeling* techniques and *multimodal fake news detection* models.

thank you
danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
spas
tack
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
gracias
хвала
asante
manana
tenki
obrigada
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
köszönöm
hvala
gracie
bayarlalaa
bedankt
enkosi
dziękuję
akun dankon
aciu
djiere dieuf
tau
mochchakkeram
дякую
mamnun
murakoze
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
dekuji
sagolun
sukriya
kop khun krap
arigatō
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
tanemirt
grazie
najis tuke
rahmet
terima kasih
kam sah hamnida
rahmat
dinlch
dhanyavadagalu
shukriya
mercé
мерси
ευχαριστώ
xièxie
তোমাকে ধন্যবাদ
감사합니다
merci