# Unmasking COVID-19 False Information on Twitter: a Topic-based Approach with BERT

Riccardo Cantini[1 [0000−0003−3053−6132]],
Cristian Cosentino[1 [0000−0002−6368−373X]],
Irene Kilanioti[2 [0000−0002−4157−3900]],
Fabrizio Marozzo[1 [0000−0001−7887−1314]], and
Domenico Talia[1 [0000−0003−1910−9236]]

[1] University of Calabria, Italy
{rcantini, ccosentino, fmarozzo, talia}@dimes.unical.it
[2] National Technical University of Athens, Greece
eirinikoilanioti@mail.ntua.gr

**Abstract.** Every day, many people use social media platforms to share information, thoughts, narratives and personal experiences. The vast volume of user-generated content offers valuable insights into the latest news and trends but also poses serious challenges due to the presence of a lot of false information. In this paper we focus on analyzing the online conversation on Twitter to identify and unveil false information related to COVID-19. To address this challenge, we devised a semi-supervised approach that combines false information detection with a neural topic modeling algorithm. By leveraging a small amount of labeled data, a BERT-based classifier is fine-tuned on the false information detection task and then is used to annotate a large amount of COVID-related tweets, organized in a topic-based clustering structure. This approach allows for effectively identifying the degree of false information in each discussion topic related to COVID-19. Specifically, our approach allows for investigating the presence of false information from a topical perspective, enabling us to examine its impact on specific topics underlying the online discussion. Among the topics with the highest incidence of false information, we found allergic reactions, microchips in vaccines, and 5G- and lockdown-related conspiracy theories. Our findings highlight the importance of leveraging social media platforms as valuable sources of information but at the same time how essential it is to identify and mitigate the impact of false information in online communities.

**Keywords:** False information · Misinformation · Disinformation · Neural Topic Modeling · COVID-19 · Natural Language Processing · BERT.

## 1 Introduction

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. With the increasing number of active users across different platforms such as Facebook, Instagram, and Twitter, social media has emerged as a vast and rich repository of valuable data [36,4]. This data, generated by billions of users worldwide, holds immense significance and potential for different fields, including business, marketing, research, and even governance [3]. The importance of social media data lies in its ability to provide real-time insights into people's thoughts, opinions, preferences, and behaviors, enabling organizations and individuals to make data-driven decisions and gain a deeper understanding of society at large.

As the influence of social media continues to grow, so does the challenge of dealing with misinformation and fake news, or more generally, false information [21]. False information is false or inaccurate information that is disseminated, either intentionally or unintentionally, leading to confusion, mistrust, and even harm. Detecting and combating false information has become a critical concern, and social media platforms play a central role in this process. For this reason, in recent years more and more researchers and companies are increasingly analyzing this phenomenon, trying to provide new solutions for detecting and mitigating the spread of false information. In this field of research, online discussions on COVID-19 represent one of the main case studies for analyzing and proposing solutions aimed at mitigating the dissemination of false information [35]. False information encompasses a wide range of topics, including vaccine efficacy and safety issues, conspiracy theories on 5G network connection, false claims on the origins of the virus, and many others, that have the potential to spread rapidly and undermine public trust in vaccination efforts. The impact of these falsehoods extends beyond the realm of social media, as they can influence individual decision-making regarding vaccine acceptance and ultimately affect public health outcomes.

In this paper we focus on analyzing the online conversation on Twitter to identify and unmask false information related to COVID-19, interpreting it from a topical viewpoint. To address this challenge, we exploited a semi-supervised approach that combines false information detection with a neural topic modeling algorithm. Our approach is divided into three main phases. Firstly, we exploit a small amount of labeled data to fine-tune a BERT-based false information detection model. Therefore, transfer learning is used to tailor the model to recognize false information in social media tweets about COVID, by adapting its pre-trained features. Subsequently, a neural topic modeling algorithm, namely BERTopic, is used to extract the main topic underlying Twitter discourse related to COVID-19, starting from a large set of unlabeled data. Lastly, we utilize the fine-tuned BERT-based classifier to determine the presence of false information for the different unlabeled posts, organized in a topic-based clustering structure through BERTopic. This process provides detailed information about the nature and extent of false information in the analyzed data, allowing us to quantitatively assess the presence of false information in the main topics discussed by users.

In contrast to state-of-the-art approaches that handle the false information problem within the large and comprehensive scope of COVID-19 discussions as a single entity [10,33,25], our approach allows for more fine-grained analysis, by taking a topical perspective. Specifically, our approach enables us to examine the impact of false information on specific topics generated during discussions. Consequently, we gain a deeper understanding of how false information influences and shapes discussions surrounding particular topics within the broader context of COVID-19. Our findings highlight the importance of leveraging social media platforms as valuable sources of information while addressing the challenges posed by false information. Furthermore, by employing a combination of false information detection and topic modeling, our work can contribute to mitigating the impact of false information in online communities.

The structure of the paper is as follows. Section 2 discusses related work in the fields of false information detection and topic modeling. Section 3 describes the devised approach. Section 4 discusses the achieved results. Finally, Section 5 concludes the paper.

## 2 Related Work

Social media plays a crucial role in information extraction and staying updated on current trends and discussions. However, the reliability of news circulating on social platforms is often questionable and susceptible to various biases. Consequently, we adopt an approach that focuses on effectively identifying topics of discussion while assessing the impact of false information on them, thus characterizing the presence of misleading and false user-generated content from a topical perspective. Therefore, our approach resides at the intersection of false information identification and topic detection. We analyze the main techniques present in the state for both research lines.

### 2.1 False information detection

With the huge amount of user-generated content on social media, assessing the reliability of online published content has become increasingly difficult in recent years. This issue derives from the presence of false information, which can come in different forms. In particular, *misinformation* refers to false information shared unintentionally, while *disinformation* implies the intentional dissemination of false or misleading information, usually for a specific purpose. Furthermore, the term *fake news* is also often used, which is a form of disinformation consisting of fabricated news aimed at deceiving public opinion.

Among the main works in the literature, addressing the detection of either misinformation or fake news, several deep learning-based approaches were proposed, leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [28,39]. Additionally, natural language processing (NLP) techniques have been increasingly used to detect false information, through the analysis of the linguistic characteristics of news articles or social media posts [30,15].

In this context, the most recent works in the literature leverage transformer-based language representation models such as BERT (Bidirectional Encoder Representation from Transformers) [7]. Such models have proven successful in a wide range of downstream tasks, by demonstrating superior performance in natural language processing and understanding. Among the main examples in the literature, FakeBERT [17] combines deep convolutional neural networks with BERT, while in [19] authors propose a combined approach that jointly leverages BERT and RNNs. In [16], a BERT-based model for fake news detection is presented, which relies on the contextual relationship between the headline and the body text of news. Furthermore, besides assessing the fake content of online news, Transformer-based architectures were also employed for fact-checking and for providing explanations. In particular, in [38] authors proposed a two-stage fake news detection system, that can both estimate the reliability of COVID-19-related claims and provide users with pertinent information about them, in the form of a textual explanation.

### 2.2 Topic detection

In recent years topic modeling has emerged as a powerful technique for uncovering latent trends and topics and extracting valuable insights from large text corpora. A wide range of topic modeling techniques have been developed, effectively applicable to a wide range of domains, such as information retrieval, document clustering, and trend detection. Among the first introduced techniques Latent Semantic Analysis (LSA) [6] uses the Singular Value Decomposition (SVD) to compute a low-rank approximation of a document term matrix (DTM) representing the corpus. LSA is simple and efficient, but it assumes a probabilistic generative model where words and documents are Gaussian distributed, which may not align with reality. To address this issue Probabilistic LSA (pLSA) was introduced, which relies on a multinomial generative model [14]. Another method is non-negative matrix factorization (NMF), it is similar to SVD but the decomposition must lead to non-negative values [23]. Latent Dirichlet Allocation (LDA) relies on the concept of mixtures of distributions to model documents as a mixture of latent topics, each of which constitutes a mixture of terms from the corpus vocabulary [18]. Among the main variants of LDA, in [1] a fuzzy version is proposed that relies on the concept of fuzzy Bag-of-Words. This fuzzy representation maps each document to a vector of keywords, where each keyword is assigned to every document with a certain membership degree. This allows for a more nuanced representation of the connections between terms and topics, accommodating the inherent ambiguity of the analyzed corpus. Another variant of LDA, which follows a deep learning approach, is LDA2Vec [27]. It mixes LDA with Word2Vec [26] by learning topic representations and latent vector representations of words simultaneously. This is achieved by modifying the standard Skip-gram model, integrating into the pivot word learnable topical information. Most recent topic modeling techniques, falling into the neural-based category, harness the power of pre-trained transformer-based Large Language Models (pLLMs) to achieve meaningful semantically-rich sentence representations. Among them,

Top2Vec [2] relies on Doc2Vec [22], while BERTopic [12] uses Sentence-BERT [31], based on siamese network architecture. Both approaches rely on the clustering of sentence-level representations projected into a low-dimensional space. Dimensionality reduction is performed using Uniform Manifold Approximation Projection (UMAP), while the HDBSCAN algorithm is used for clustering. Finally, topic representations are extracted from the topic-based clustering structure by selecting the nearest neighbors of the cluster centroid, in the case of Top2VEc, and by applying a class-based tf-idf, in the case of BERTopic.

## 3 Proposed Approach

This work focuses on the analysis of user-generated content on Twitter to identify and investigate false information related to COVID-19. For this purpose, we devised a semi-supervised approach that leverages a combination of false information detection and topic modeling, to achieve a topic-oriented representation of false information. Specifically, a BERT classifier is fine-tuned on a small set of annotated data, to make it able to identify false information present in a given post. Then, unlabeled data are used to unveil the main COVID-related topics of discussion underlying social media conversation. Specifically, this step relies on BERTopic, one of the most used neural topic modeling methods in the literature, which leverages semantically-rich sentence representations achieved through pre-trained LLMs. Finally, a false information score is computed for each topic identified by BERTopic, through the use of the fine-tuned false information detection model. This process allows for a topic-oriented quantification of the impact of false information on Twitter conversations about COVID-19. Hence, by following this approach, we can highlight the main discussion topics that are most affected by false information, from a quantitative viewpoint, while also finding concrete examples of misinformed user-generated content related to these topics. In the following, we provide a detailed description of the main steps of our approach, whose execution flow is depicted in Figure 1.

### 3.1 Fine-tuning of the false information detection model

In this step, a BERT model is fine-tuned for the false information detection task. Specifically, starting from a small set of labeled posts, we train a binary classifier to detect false information using a transfer learning approach. Indeed, Large Language Models (LLMs) like BERT have proven successful in a wide range of downstream tasks, through the adaptation of pre-trained features to specific purposes. Besides BERT, other optimized variants exist, each introducing improvements in both the architecture and the pre-training phase. Due to this, we tested several BERT-like models, including BERT, ALBERT, BERTWEET, DISTILBERT, and ROBERTA, to find out the best trade-off between classification accuracy and training/inference times. The BERT model as well as the evaluated variants was fine-tuned for detecting false information in social media posts. Specifically, during this step, pre-trained weights are slightly adapted to

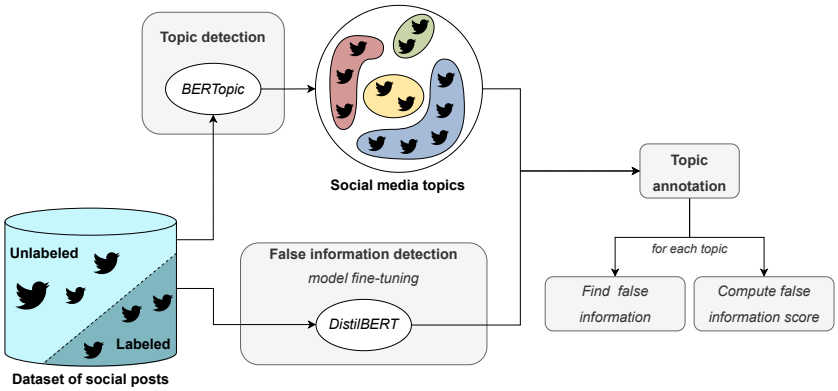



**Fig. 1.** Execution flow of the proposed approach.

the binary downstream task under consideration, by using a binary cross-entropy loss, the ADAM optimization algorithm, and a small learning rate, which is crucial to correctly transfer knowledge from BERT by avoiding pre-trained weights to be distorted by large weight updates.

### 3.2 Topic detection

In this step, starting from a large set of unlabeled posts, we use topic modeling to extract the main discussion topics underlying social media conversation. To this purpose, we leveraged BERTopic, a neural topic modeling technique that relies on Transformer-based pLLMs to generate semantically-rich vector representations of the sentences in a corpus. As recently demonstrated in the literature, the use of neural approaches like BERTopic for topic modeling leads to superior performance in terms of coherence and diversity [12,11,9]. In particular, in BERTopic Sentence-BERT is utilized for sentence embedding, which uses siamese neural network structures to generate semantically meaningful and comparable sentence representations. Then, such representations are projected in a low-dimensional space using UMAP (Uniform Manifold Approximation and Projection), and clustered into semantically-related groups via HDBSCAN. Following this approach, BERTopic can identify a topic-based clustering structure from which topic representations are computed, one for each cluster, using a class-based version of tf-idf.

### 3.3 Topic annotation

Our approach adopts a topic-oriented perspective to thoroughly analyze the impact of false information within social media conversations. Therefore, in this step, we identify the discussion topics that are most affected by false information and quantify the extent of false information prevalent within them.

Specifically, the false information detection model fine-tuned previously is used to determine a false information probability for each unlabeled sentence. Thus, given a cluster, i.e., a topic, a false information score $S(c)$ associated with that topic is computed as follows.

$$S(c) = \frac{\sum_{s \in c} p_s^c \;\; p_s^{fi}}{\sum_{s \in c} p_s^c}, \quad \text{where } c \; 2 \; C \tag{1}$$

In the above formula, $p_s^c$ indicates the degree of membership of sentence $s$ to the cluster $c$, while $p_s^{fi}$ is the sigmoid output of BERT, which specifies a soft-label for the sentence $s$ measuring its degree of false information. Therefore, the false information score $S(c)$ for cluster $c$ is determined as the average false information of the sentence contained in that cluster, weighted on the probability of those sentences.

## 4 Experimental Results

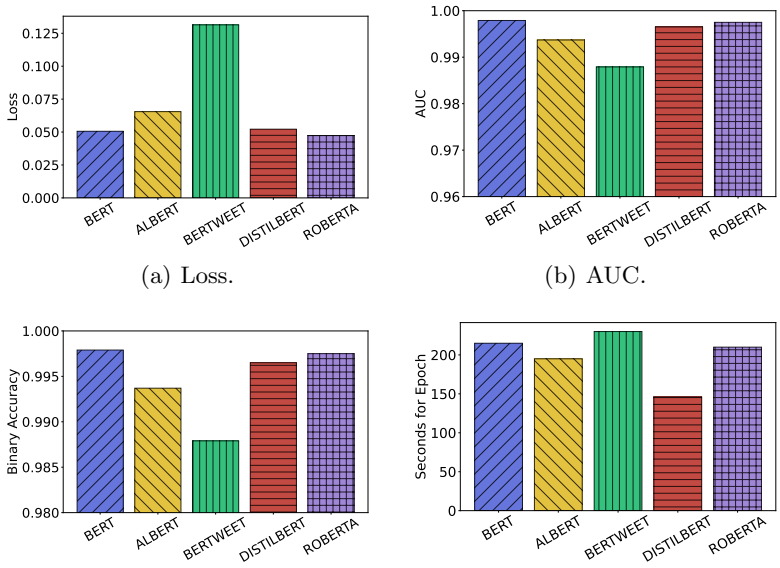
The COVID-19 pandemic has not only had a profound impact on society but has also led to the widespread dissemination of false information on social media, resulting in increased vaccine hesitancy and the proliferation of conspiracy theories. Therefore, as stated in Section 1, the goal of this work is to detect the main false information present in COVID-related discussions, characterizing it from a topical perspective. To this purpose, we applied our approach to the *ANTi-Vax* dataset [13], composed of tweets from December 1, 2020, until July 31, 2021 related to the COVID-19 (SARS-CoV-2) pandemic. The dataset consists of a small portion of labeled data (about 15K) and a large set of unlabeled data (about 15M). Labeled data was manually annotated and validated by health medical experts, into two classes: *false information* for all those tweets that contain common myths and misinformation (e.g., the vaccine contains tracking device), or *reliable content*. It must be noted that all sarcastic and humorous tweets have not been included as false information. Among all unlabeled data, we focused on posts generated in the month of January 2021, encompassing 303,541 tweets. In the following, the experimental results we achieved will be comprehensively discussed, focusing on: (*i*) the choice of the best-suited transformer-based model to be fine-tuned for the false information detection task; (*ii*) the main identified topics that drove COVID-related discourse on Twitter; (*iii*) the analysis of COVID-related false information from a topical perspective.

### 4.1 Model selection for false information detection

Among the main alternative models that can be effectively used for the binary task of false information detection, describe in detail in Section 2.1, we choose to follow a transfer-learning approach, by fine-tuning a BERT-based classifier on our downstream task. To select the most suitable model for our purposes, we conducted a comparative analysis of the following models.

- **BERT**: it is a pre-trained language representation model based on the transformer architecture [7].
- **ALBERT**: it is a lightweight variant of BERT. It introduces some improvements such as factorized embedding parameterization, and inter-sentence coherence loss, by replacing the Next Sentence Prediction with the Sentence order prediction task during pre-training [20].
- **BERTWEET**: it is a Twitter-specific variant of BERT, trained on Twitter text data. It manages unique Twitter features such as hashtags, mentions, URLs, and emojis [29].
- **DISTILBERT**: it is a distilled version of BERT, with about 40% fewer parameters. This reduction in size, achieved through a knowledge distillation approach, allows for faster training, making the model less resource-intensive [34].
- **ROBERTA**: it is an improved version of BERT, which removes the Next Sentence Prediction (NSP) task from the pre-training phase and introduces dynamic masking to vary the masked tokens during language modeling [24].

The performance evaluation of the different BERT-like models (i.e., BERT, ALBERT, BERTWEET, DISTILBERT, ROBERTA) was conducted on a held-out test set, encompassing 3000 samples, considering four metrics: score loss, AUC (Area Under the Curve), binary accuracy, and training time (measured in seconds per epoch). Figure 2 shows the scores obtained from different models for each considered metric.




(a) Loss. (b) AUC.

(c) Binary Accuracy. (d) Seconds for epoch.

**Fig. 2.** BERT-based model comparison for false information detection.

Figure 2(a) shows the loss values achieved by each model, computed with a binary cross-entropy. Models such as BERT, DISTILBERT, and ROBERTA demonstrate lower loss values, indicating a better fit to the test data. Figure 2(b) and Figure 2(c) display the Area Under the Curve (AUC) values and the binary accuracy for each model, which measure the model's ability to correctly distinguish between negative and positive classes. Also in this case, we observe that BERT, DISTILBERT, and ROBERTA exhibit the highest values. Finally, Figure 2(d) shows the time required by each model to complete an epoch during training. The DISTILBERT model stands out as the fastest among the compared models, showcasing its capability for fast and efficient training. All achieved results are summarized in Table 1, showing the average values obtained from multiple experiments, which exhibit a negligible variance. Summing up, what emerges from our evaluation is that the DISTILBERT model achieves the best trade-off between accuracy and training time. Consequently, we utilized the DISTILBERT model as the reference model for false information classification throughout all the subsequent experiments.

| Model | Version | Loss | AUC | Binary accuracy | Seconds per epoch |
|---|---|---|---|---|---|
| **BERT** | bert-base-uncased | 0.050 | 0,998 | 0.982 | 215 |
| **ALBERT** | albert-base-v2 | 0.066 | 0.994 | 0.980 | 195 |
| **BERTWEET** | bertweet-base | 0.131 | 0.988 | 0.957 | 230 |
| **DISTILBERT** | distilbert-base-uncased | 0.052 | 0.997 | 0.986 | 146 |
| **ROBERTA** | roberta-base | 0.047 | 0.997 | 0.983 | 210 |

**Table 1.** BERT-based model comparison for false information detection.

### 4.2 COVID-related detected topics

The topic detection phase, as described in Section 3.2, relies on BERTopic, which has proven effective in identifying discussion topics in social media data [12,11]. In our experiments, the application of BERTopic led to the identification of several topics that shaped the COVID-related discussion on Twitter. These topics will be used in the final step, to characterize the identified false information from a topical perspective.

Among the main identified topics, within the broader topic of *COVID vaccines* we found the discussion about the efforts and strategies of the US president *Joe Biden* and the former UK prime minister *B. Johnson*. Additionally, the online conversation focused on specific vaccines such as the *P zer vaccine* and *Johnson & Johnson*, discussing their effectiveness and side effects, such as *allergic reactions* and the risks related to *pregnancy and breastfeeding*. The Twitter discourse was also centered on the *European Union*'s approach to managing the pandemic and anti-contagion rules such as *mask wearing* and *lockdown*, also debating the effects on major sporting events like the *Olympic Games and NBA*.

Furthermore, users discussed the long-term effects of COVID, especially on *older individuals*, and other conspiracy theories about the presence of *microchip* inside vaccines. Other identified topics include *Dr. A. Fauci*, *vaccine passports*, the impact of *COVID-19 in Florida*, *school-related issues*, and the challenges faced by *workers and employers*.

To evaluate the identified topics we used Topic coherence and diversity. Coherence measures how closely related and meaningful are the words within a topic, thus giving an estimate of how well they express a specific theme or concept. Among the main coherence metrics, we used CV [32] and Normalized Pointwise Mutual Information (NPMI) [5], achieving a value of 0.51 and 0.09 respectively. Differently, topic diversity assesses how different and unrelated the topics are from each other, which is necessary to comprehensively represent the corpus. We used the Percentage of Unique Words (PUW) [8] and the average pairwise Jaccard Distance (JD) [37], achieving a value of 0.97 and 0.99 respectively. Similarly to the experimental evaluation present in [12], we computed each metric by averaging across 10 different runs. In addition, for each run, metrics are averaged by varying the number of topics from 10 to 50, with steps of 10.

### 4.3 Topic-oriented false information detected in COVID discussions

In contrast to state-of-the-art approaches that treat the false information problem within the large and comprehensive scope of COVID-19 discussions as a single entity, our approach allows for more granular analysis. Specifically, our approach enables us to examine the impact of false information on the Twitter discourse from a topical perspective. Figure 3 shows the discussion topics ordered according to the level of false information present in them. Specifically, for each topic, we computed a false information score, i.e., $S(c)$ as defined in Section 3.3, which quantifies the extent of false information present in it.

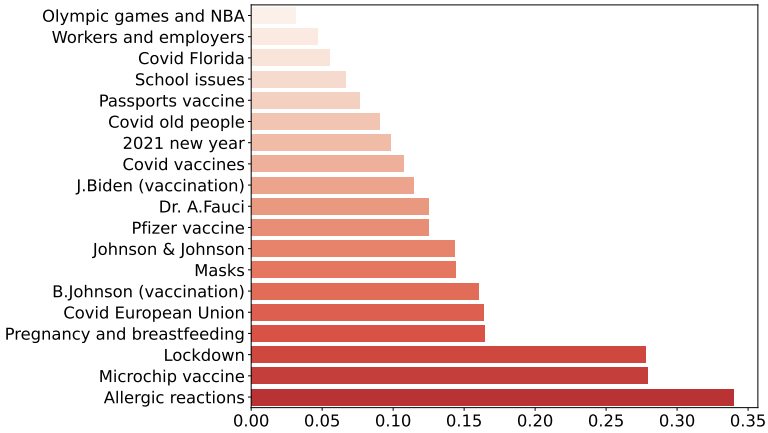
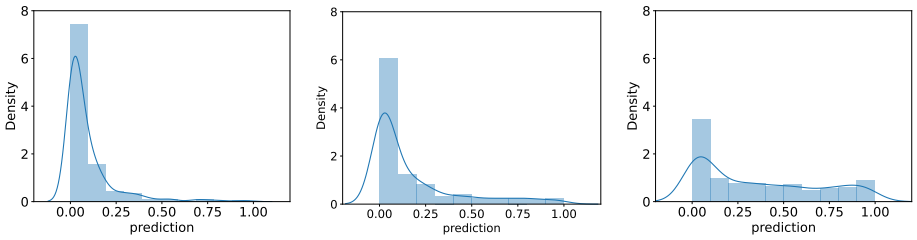



**Fig. 3.** False information score for each identified topic.

From Figure 3 it can be observed that there is a varying range of false information levels across the different topics. In the following we report three of these topics, characterized by the minimum, maximum, and median value of false information, according to the distribution of the $S(c)$ score.

- *Olympic games and NBA*: this topic refers to the Tokyo Olympics, held in Japan, and specifically to Olympic athletes and NBA basketball players.
- *Dr. A. Fauci*: this topic refers to Dr. Anthony Fauci, a renowned infectious disease expert in the United States, addressing conspiracy theories and spreading scientific information about vaccinations.
- *Allergic reactions*: this topic refers to severe side effects and physical symptoms that may occur after receiving a vaccine injection.

For each of the highlighted topics, we also show the distribution of the output achieved by the fine-tuned DistilBERT classifier. This model, as described in Section 3, computes a probability value, i.e. $p_s^{fi}$, indicating how likely it is that a given content is false information. Therefore, given a sentence $s$ a value of $p_s^{fi}$ close to 0 indicates a low probability that this sentence contains false information, while values close to 1 represent the opposite case. The achieved results, shown in Figure 4, are in line with the false information scores computed previously. In particular, Figure 4(a) referred to the Olympic Games and NBA, shows a distribution whose values are predominantly concentrated toward 0, indicating a higher prevalence of non-false information predictions. A similar unimodal distribution is achieved in Figure 4(b), related to the topic of Dr. A. Fauci. Differently, by observing Figure 4(c), related to the topic with the highest false information score (i.e., allergic reactions), a bimodal distribution clearly emerges, indicating a non-negligible presence of predictions very close to 1. This translates into a greater presence of user-generated content identified by DistilBERT as false information, mainly related to untested hypotheses about serious health side effects caused by vaccines.



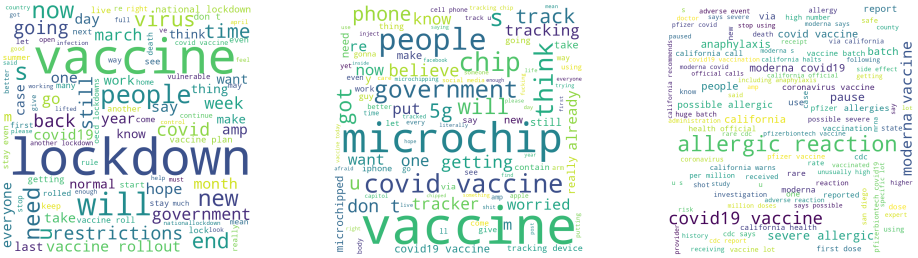

(a) Olympic games and NBA. (b) Dr. A. Fauci. (c) Allergic reactions.

**Fig. 4.** Examples of distribution $p_s^{fi}$ of topics with low, medium and high levels of false information score.

As a final analysis, we focused on the top-3 topics with the highest misinformation score:

- *Allergic reactions*, already described above. False information score: 0.35.
- *Microchip vaccine*: the topic refers to conspiracy theories regarding the presence of microchips in vaccines. False information score: 0.25.
- *Lockdown*: the topic refers to lockdown measures imposed during the COVID-19 pandemic. False information score: 0.24.

For enriching the description of these topics, Figure 5 presents the word clouds referring to them, which allows for highlighting their main keywords and concepts in a graphical manner. Among the most significant words we found: $(i)$ *allergic reaction, vaccine, severe allergic, modern*, and *anaphylaxis* for the allergic reaction topic; $(ii)$ *microchip, vaccine, 5g*, and *track* for the microchip topic; $(iii)$ *lockdown, vaccine, government*, and *restriction* for the lockdown topic.




(a) Lockdown. (b) Microchip vaccine. (c) Allergic reactions.

**Fig. 5.** Word cloud representations of the top-3 topics per false information.

For the sake of completeness, in Table 2 we report an example of detected false information, for each of the three topics. For each example, we indicate the related topic, the text of the tweet, the false information probability given by DistilBERT ($p_s^{fi}$), and the degree of membership to the assigned cluster ($p_s^c$).

| 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 |

**Table 2.** Example tweets, identified as false information, for each of the top-3 topics per false information score.

The reported tweets express (*i*) skepticism and concerns about lockdown measures and COVID-19 vaccines; (*ii*) conspiracy theories, including the belief that vaccines emit harmful 5g waves and contain surveillance microchips; (*iii*) serious side effects of vaccines and no-vax instigations.

## 5 Conclusion

Social media has revolutionized the way we communicate and share information, providing valuable data with high potential for many fields of application. However, alongside these benefits, there has been an alarming surge in the proliferation of false information and fake news, necessitating urgent measures to mitigate their impact.

This paper focuses on the analysis of Twitter conversations to uncover and address false information pertaining to COVID-19. Employing a semi-supervised strategy and harnessing the capabilities of a BERT-based classifier, the study effectively identifies and annotates different topics present in online conversations, while evaluating the extent of false information associated with each topic. These encompass allergic reactions, microchips in vaccines, 5G conspiracy theories, and the impact of lockdown measures.

In contrast to state-of-the-art approaches that treat the false information problem within the large and comprehensive scope of COVID-19 discussions as a single entity, our approach allows for a finer-grained analysis, enabling us to examine the impact of false information on specific topics generated during discussions. Through the employment of transfer learning for false information detection and neural topic modeling, our work not only aids in identifying specific instances of false information but also provides insights into the underlying factors and dynamics contributing to its spread. This understanding is crucial for developing targeted interventions and strategies that effectively combat the dissemination of false information, ultimately strengthening the reliability and trustworthiness of information shared on social media platforms.

## Acknowledgements

This work has been partially supported by by the "National Centre for HPC, Big Data and Quantum Computing", CN00000013 - CUP H23C22000360005, and by the "FAIR – Future Artificial Intelligence Research" project - CUP H23C22000860006.

## References

1. Akhtar, N., Sufyan Beg, M., Javed, H.: Topic modelling with fuzzy document representation. In: Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3. pp. 577–587. Springer (2019)

2. Angelov, D.: Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470 (2020)
3. Belcastro, L., Cantini, R., Marozzo, F.: Knowledge discovery from large amounts of social media data. Applied Sciences **12**(3) (2022)
4. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Learning political polarization on social media using neural networks. IEEE Access **8**, 47177–47187 (2020)
5. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL **30**, 31–40 (2009)
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics **8**, 439–453 (2020)
9. Egger, R., Yu, J.: A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in sociology **7** (2022)
10. Enders, A.M., Uscinski, J.E., Klofstad, C., Stoler, J.: The different forms of covid-19 misinformation and their consequences. Harvard Kennedy School Misinformation Review (2020)
11. Gabarron, E., Dorronzoro, E., Reichenpfader, D., Denecke, K.: What do autistic people discuss on twitter? an approach using bertopic modelling (2023)
12. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
13. Hayawi, K., Shahriar, S., Serhani, M.A., Taleb, I., Mathew, S.S.: Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. Public health **203**, 23–30 (2022)
14. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57 (1999)
15. Jarrahi, A., Safari, L.: Evaluating the effectiveness of publishers' features in fake news detection on social media. Multimedia Tools and Applications **82**(2), 2913–2939 (2023)
16. Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H.: exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Applied Sciences **9**(19), 4062 (2019)
17. Kaliyar, R.K., Goswami, A., Narang, P.: Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia tools and applications **80**(8), 11765–11788 (2021)
18. Korshunova, I., Xiong, H., Fedoryszak, M., Theis, L.: Discriminative topic modeling with logistic lda. Advances in neural information processing systems **32** (2019)
19. Kula, S., Choraś, M., Kozik, R.: Application of the bert-based architecture in fake news detection. In: 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12. pp. 239–249. Springer (2021)
20. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)

21. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
22. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196. PMLR (2014)
23. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
25. Loomba, S., de Figueiredo, A., Piatek, S.J., de Graaf, K., Larson, H.J.: Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. Nature human behaviour **5**(3), 337–348 (2021)
26. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
27. Moody, C.E.: Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019 (2016)
28. Nasir, J.A., Khan, O.S., Varlamis, I.: Fake news detection: A hybrid cnn-rnn based deep learning approach. International Journal of Information Management Data Insights **1**(1), 100007 (2021)
29. Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200 (2020)
30. de Oliveira, N.R., Pisa, P.S., Lopez, M.A., de Medeiros, D.S.V., Mattos, D.M.: Identifying fake news on social networks based on natural language processing: trends and challenges. Information **12**(1), 38 (2021)
31. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
32. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. pp. 399–408 (2015)
33. Roozenbeek, J., Schneider, C.R., Dryhurst, S., Kerr, J., Freeman, A.L., Recchia, G., Van Der Bles, A.M., Van Der Linden, S.: Susceptibility to misinformation about covid-19 around the world. Royal Society open science **7**(10), 201199 (2020)
34. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
35. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data **8**(3), 171–188 (2020)
36. Talia, D., Trunfio, P., Marozzo, F.: Data Analysis in the Cloud: Models, Techniques and Applications. Elsevier (October 2015), iSBN 978-0-12-802881-0
37. Tran, N.K., Zerr, S., Bischoff, K., Niederée, C., Krestel, R.: Topic cropping: Leveraging latent topics for the analysis of small corpora. In: Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3. pp. 297–308. Springer (2013)
38. Vijjali, R., Potluri, P., Kumar, S., Teki, S.: Two stage transformer model for covid-19 fake news detection and fact checking. arXiv preprint arXiv:2011.13253 (2020)
39. Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S.: Ti-cnn: Convolutional neural networks for fake news detection. arXiv preprint arXiv:1806.00749 (2018)