Tackling the COVID-19 Conspiracies: The Data-Driven Approach

Nenad Petrović

Abstract - **Apart from obvious impact on many aspects of our everyday lives, the current COVID-19 pandemic has led to occurrence many conspiracy theories as side-effect, especially via social media. However, in such confusion among population, the situation becomes even more difficult to handle, which could potentially have even more dramatic consequences on global economy and increase the number of lives lost. In this paper, a data-driven framework for analysis of the facts related to COVID-19 disease is proposed. The proposed implementation leverages semantic knowledge representation, fuzzy reasoning and correlation analysis.**

Keywords - **Coronavirus, Correlation analysis, Fuzzy reasoning, Ontology, Semantic knowledge representation.**

I. Introduction

Since the beginning of this year, the pandemic of novel infectious disease *COVID-19* has affected almost all aspects of our everyday lives. It was first discovered in China, but spread quickly to other continents in just few weeks [1-3]. According to [4], until June 15th, 2020, the total number of identified cases was 8,039,649, while the disease has taken 436,498 lives worldwide. However, due to huge number of cases in some countries, several measures were taken by governments in order to reduce to disease spread, such as social distancing, limiting citizens' movement within country boarders and abroad, often together with prohibition and cancellation of huge public events [5]. Therefore, from work to personal relations and entertainment, COVID-19 pandemic has brought many changes to our daily routine, habits and activities. Despite the fact that the pandemic seems weaker nowadays, some of the government measures are still applied in some countries due to unstable situation.

However, one of the crucial problems related to COVID-19 which led to dramatic consequences are the speed of its spread and lack of approved vaccine and medication [6]. for the disease. Apart from huge number of lost lives and stagnation of economy as well [7], the situation becomes even more difficult due to spread and circulation of conspiracy theories worldwide [8, 9], especially via social networks [10]. One of the most common conspiracy theories is the relation between the deployment of 5G network and COVID-19 spread [8-10], which even lead to organized vandalism against 5G equipment [11].

In this paper, a data-driven approach is adopted to tackle the conspiracies in context of COVID-19 disease, leveraging fuzzy reasoning, performed on facts stored within the semantic knowledge base created as a result of ontology learning against open scientific literature and making use of correlation analysis of public data sets for truth value determination. To the best of author's knowledge, such solution has not been publicly presented yet or documented in published scientific literature.

II. Background and Related Work

A. Fuzzy Logic

Fuzzy logic refers to a form of multi-valued logic where the truth values, apart from 0 and 1 could be any real number between 0 and 1. Its goal is to handle the concept of partial truth, where the truth value may range between completely true and completely false. It is similar to human reasoning as it relies on imprecise information to make decisions. Traditional logic which requires deeper understanding of a system, such as exact equations and precise numeric values. On the other side, fuzzy logic enables modeling complex systems using a higher level of abstraction originating from human knowledge and experience [12, 13]. In many cases, control of complex systems by experienced human operators is more successful and effective than automatic methods. Fuzzy reasoning has been adopted in many complex systems that cannot be modelled precisely even under various assumptions and approximations incorporating the human expert knowledge [13].

The medical decision-making process is fuzzy in its nature [14]. During this process, the doctors need to reason about linguistic concepts in order to bring a decision about diagnosis or make prognosis. However, the conversion from fuzzy nature into outcome often leads to loss of precision [14], which could have dramatic consequences, when it comes to patient treatment. For that reason, fuzzy logic is a suitable way to provide the doctors with the support needed for handling of linguistic concepts and avoiding such loss of precision. Fuzzy logic technologies are applied across many areas of healthcare, and they have been approved as quite effective [14, 15]. For example, in [16], a fuzzy medical diagnosis expert system based on quantiles of diagnostic measures was presented, while the work presented in [17] focuses on coronary heart disease risk determination based on cholesterol measurements.

Nenad Petrović is with University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, E-mail: nenad.petrovic@elfak.ni.ac.rs

In this paper, fuzzy reasoning is applied in wider context, in order to take into account the partial truth of novel facts about COVID-19 available online within the decision-making process. However, the proposed approach can be also applied to support medical decision-making when it comes to COVID-19 patient assessment and treatment in order to eliminate any possible conspiracies that can have impact on treatment.

B. Correlation Analysis

Correlation analysis refers to a statistical method which is used for discovery of a relationship between two variables and how strong that relationship is [18]. Correlation coefficient represents a numerical measure of correlation between the relative movements of two variables from the observed data set. Its value is in range between -1.0 and 1.0. Positive value of correlation means that for positive increase in one variable, there is also a positive increase in the second variable. On the other side, negative correlation value indicates that the variables move in opposite directions - for a positive increase of one variable, the second variable decreases. In case of correlation value which is exactly 0, there is no linear relationship between the observed variables. There are many types of different correlation coefficients, but the most commonly used is the Pearson correlation coefficient (denoted as *r*) that measures the strength and direction of the linear relationship between two variables [18]. However, it is not able to capture nonlinear relationships between two variables and distinguish between dependent and independent variables.

There are several works that adopt correlation coefficient to discover relationships between some factors related to COVID-19 pandemic. In [19], the analysis of the correlation between confirmed cases of COVID-19 in Spain and several geographic, meteorological and socioeconomic variables at the province level was presented. On the other side, in [20], a correlation between racial demographics and COVID-19 cases was analyzed.

In this paper, MATLAB's implementation of Pearson correlation known as *corcoeff*[1] was used to calculate the value of correlation between two variables relevant to COVID-19 based on publicly available online data sets. Moreover, the calculated value is used as truth value for facts stating that these two observations are related, which is also leveraged within fuzzy reasoning process relying on semantic representation of these facts based on ontologies.

C. Semantic Knowledge Representation

Semantic technology enables encoding the meaning of data separately from the content itself and related applications, providing the ability to understand data, exchange that understanding and perform reasoning on top of it. The formalization of this knowledge representation is made in a form understandable and suitable for use by both humans and computers, which is an advantage for usage in reasoning-based systems. Within the semantic knowledge bases, the data is represented with respect to ontologies. An ontology consists of classes, individuals, attributes and relations. Classes represent abstract groups, collections or types of objects. Individuals are instances of these classes. Attributes refer to properties, characteristics and parameters of classes. Relations define ways in which classes and individuals are related. For both the ontologies and facts, the RDF standard language is used. It is used to define (subject, predicate, object) triplets that are persisted within the semantic triple stores. On the other side, SPARQL is used for query execution against the semantic triple stores. The results are retrieved in order to support different reasoning mechanisms that enable inference of new knowledge and facts from the existing knowledge base.

In [21], ontologies were used for representation of pandemic simulation scenarios in case of H5N1. On the other side, in [22], Epidemiology Ontology was introduced to enable sharing of epidemiology resources within semantic-enabled platforms. Recently, in [23], an extension of the Infectious Disease Ontology was introduced, including the concepts relevant to coronavirus pandemic.

In this paper, semantic knowledge representation is used for facts about coronavirus pandemic coming from various online textual sources - open-access scientific paper repositories (such as PubMed[2]) and web pages. The transformation of textual data to semantic representation is performed relying on ontology learning module built upon the work presented in [24]. Moreover, an ontology is also used for representation of fuzzy rules used by the mechanism of reasoning itself, in a similar way as it was done in [25] to support explainable artificial intelligence based on fuzzy reasoning in smart cities.

III. Implementation Overview

In the first step, the user specifies the desired hypothesis about COVID-19 disease that will be tested using a graphical user interface in web browser. The tested hypothesis is given either in form of *(condition set->consequence)* or pair of potentially connected concepts that are in relation $r(concept_1, concept_2)$. A condition set consists of a set of statements connected with *and* - $(condition_1 \wedge condition_2 \wedge \ldots condition_n)$. After that, the key terms are extracted from provided user-defined hypothesis elements ($term_1$ is extracted from $concept_1$ or $term_2$ is extracted from $condition_2$, for example) and forwarded as input to the search and retrieval module. This module has role to execute search queries for key terms against repositories of publicly available scientific texts or extract numerical statistic data related to COVID-19 pandemic from web pages. Furthermore, the retrieved raw data is filtered and stored into a form suitable for further steps. Moreover, the prepared data is used as input for two different processing modules. While textual data is forwarded to ontology learning module to construct a semantic knowledge base about given terms, the numerical data is leveraged for correlation analysis by calculating the correlation coefficient between the data about input pair of concepts.

[1] https://www.mathworks.com/help/matlab/ref/corrcoef.html

[2] https://pubmed.ncbi.nlm.nih.gov/

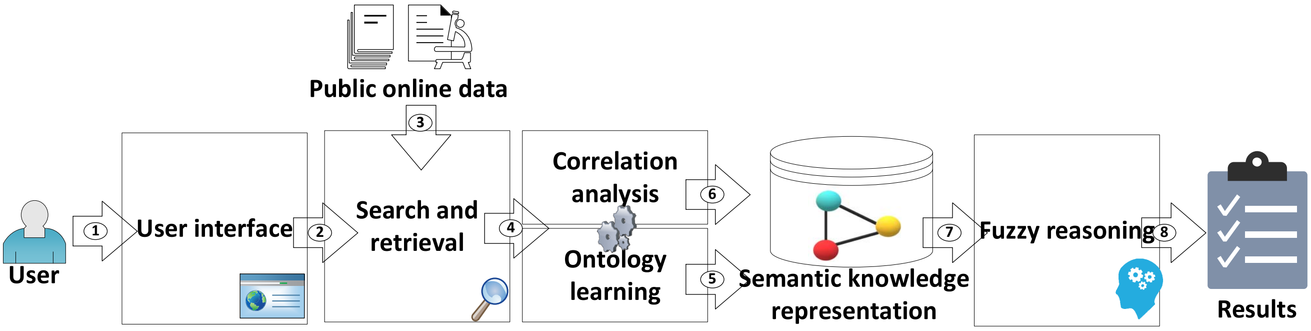

Fig. 1. Framework overview: 1-User-defined hypothesis 2-Key words for terms from hypothesis 3-Raw textual or numerical data about key terms 4-Data prepared for processing 5-Representation of facts extracted from text in semantic triplet form 6-Correlation coefficient values 7-Facts for given terms found in semantic knowledge base 8-Interpretation of results obtained as output of fuzzy reasoning process

The truth value of facts stored within the semantic knowledge base takes a real number value from range [0, 1] and is calculated in case of concept relation testing according to formula:

$$T_{t_1,t_2} = \frac{\frac{n_{(t_1,t_2)}}{n_{texts}} + correlation\_coeff(t_1,t_2)}{2} \tag{1}$$

In Eq. (1), $n_{(t1,t2)}$ denotes the number of texts where the relation between the terms extracted from concepts given in the hypothesis is present; n_{texts} is the total number of texts returned for given terms; *correlation_coeff(t1, t2)* refers to correlation coefficient between the variables relevant to these terms extracted from public datasets. In case that search module fails to find the relevant datasets for correlation analysis, than the truth value is not calculated as average, but only taking into account the percentage of texts where the relation from the hypothesis is present ($n_{(t1,t2)}/\ n_{texts}$), instead. On the other side, in formula for the other type of hypotheses considered, t_1 represents a condition from the condition set, while t_2 refers to the consequence. Moreover, the relevance of public online data can be taken into account in order to reduce the impact of low-quality publications by including a trustworthiness coefficient of data source where the considered fact was collected, denoted as $trust_i$:

$$\frac{\sum_{i=1}^{N} trust_i \cdot n_{(t_1,t_2)}}{n_{texts}} \tag{2}$$

For this value, either impact factor of scientific publication can be taken into account or some other metric, such as page rank or relevance [26]. Once the truth values for the collected facts used are ready, the fuzzy reasoning process can begin. It executes a set of SPARQL queries against the semantic knowledge base to check whether the tested conditions exist and retrieves their truth values. However, after that, there are two possible flows, depending on type of hypothesis testing. In case of concept relation testing, if the truth value of facts describing the assumed relationship between the concepts *c1* and *c2* is below given threshold $threshold_{conspiracy}$ (a real-valued number between 0 and 1), than the tested hypothesis is considered as a conspiracy. On the other side, when it comes to the other type of hypotheses in form of *condition set->consequence*, then the hypothesis passes the conspiracy test only in case when all the facts that represent the relation between a $condition_i$ from the condition set and *consequence* have truth value above $threshold_{conspiracy}$. Finally, the results achieved as outcome of the described reasoning process are interpreted (whether hypothesis is considered as a conspiracy) and shown to the user via GUI.

In Fig. 1, an overview of the proposed framework is summarized. Moreover, in Listing I, pseudocode of the conspiracy test based on fuzzy reasoning is given.

```
Input: hypothesis, $threshold_{conspiracy}$
Output: true/false
Steps:
Extract key terms from hypothesis;
For each term t in set of key terms
    Retrieve texts from online sources about term t;
    term facts:=Perform ontology learning about term t based on texts;
    Add term facts to the collection of all facts;
    Retrieve data set related to term t if possible;
    Discover transitive relations;
    Calculate truth based on (1);
End for each;
conspiracy:=false;
For each fact f in set of facts
    If (f.truth < $threshold_{conspiracy}$)
        conspiracy:=true;
    End if;
End for each;
return not(conspiracy);
End
```

Listing 1. Pseudocode of COVID-19 conspiracy testing algorithm

IV. Experiments and Evaluation

For evaluation, a laptop equipped with Intel i7 7700-HQ quad-core CPU running at 2.80GHz, equipped with 16GB of DDR4 RAM and 1TB HDD. In Table 1, the processing times of various steps achieved in several experiments of hypothesis testing are given. PubMed papers were used as basis for textual corpus by the ontology learning module, while the numerical statistics used for correlation relied mostly on [4], together with several pages from [27]. The value of $threshold_{conspiracy}$ was 0.8 in all cases.

The presented framework was able to correctly distinguish between conspiracy (increased investment in 5G increases COVID-19 cases) and valid scientific results (death rate is higher among elderly and persons with chronic diseases). As it can be noticed, the processing time increases with textual corpus size and data set size in tests where correlation analysis was performed, as expected.

TABLE I
EXPERIMENT RESULTS

H	N_{texts}	$T_{ontology}$ [s]	T_{correl} [s]	T_{reason} [s]	Out
increase (5G, cases)	10	98	0.13	0.86	false
old->deaths	20	147	0.25	0.91	true
chronic->deaths	30	289	-	1.01	true

V. CONCLUSION AND FUTURE WORK

With respect to the initial results, it can be recognized that the presented approach has huge potential for adoption within various systems and use cases. In future, it is planned to integrate it within medical expert systems based on fuzzy reasoning, such as the one presented in [17], with purpose of decisioning improvement, especially when it comes to impact of COVID-19 on patients with chronic heart diseases. Finally, the presented solution will be integrated within the framework for efficient resource planning in pandemic crisis with goal of patient risk assessment [28] and leveraged by mobile app [29].

ACKNOWLEDGEMENT

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] G. Spiteri et al., “First Cases of Coronavirus Disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020”, Eurosurveillance, pp. 1-6, 2020.

[2] “Cluster of Pneumonia Cases Caused by a Novel Coronavirus, Wuhan, China” [online], European Centre for Disease Prevention and Control, Stockholm, pp. 1-10, 2020. https://www.ecdc.europa.eu/sites/default/files/documents/Risk%20assessment%20-%20pneumonia%20Wuhan%20China%2017%20Jan%202020.pdf

[3] B. Cruz, M. Dias, “COVID-19: From Outbreak to Pandemic”, GSJ, vol. 8, no. 3, March 2020, pp. 2230-2238.

[4] Coronavirus Update (Live) [online]. Available on: https://www.worldometers.info/coronavirus/

[5] Considerations Relating to Social Distancing Measures in Response to COVID-19 – Second Update [online]. Available: https://www.ecdc.europa.eu/sites/default/files/documents/covid19-social-distancing-measuresg-guide-second-update.pdf

[6] Y. Song et al., “COVID-19 Treatment: Close to a Cure? – A Rapid Review of Pharmacotherapies for the Novel Coronavirus” [preprint], pp. 1-25, 2020, https://doi.org/10.20944/preprints202003.0378.v1

[7] P. Vanini, “Protection of the Population and the Economy in a Pandemic” [preprint], pp. 1-21, 2020.

[8] D. Freeman et al., “Coronavirus Conspiracy Beliefs, Mistrust, and Compliance with Government Guidelines in England”, Psychological Medicine, pp. 1–13, 2020. https://doi.org/10.1017/S0033291720001890

[9] M. Uthman et al., “5G Radiation and COVID-19: The Non-Existent Connection”, International Journal of Research im Electronics and Computer Engineering, vol. 8, no. 2, April-June 2020, pp. 34-38.

[10] W. Ahmed et al., “COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data”, J Med Internet Res 2020; 22(5):e19458, pp. 1-9, 2020.

[11] Cell Tower Vandals and Re-open Protestors — Why Some People Believe in Coronavirus Conspiracies [online]. Available on: https://theconversation.com/cell-tower-vandals-and-re-open-protestors-why-some-people-believe-in-coronavirus-conspiracies-138192

[12] S. G. Tzafestas, N. Venetsanopoulos et al., “Fuzzy Reasoning in Information, Decision and Control Systems”, Kluwer Academic Publishers, 1994, https://doi.org/10.1007/978-0-585-34652-6

[13] C. Carlsson, R. Fullér, “Fuzzy Reasoning in Decision Making and Optimization. Studies in Fuzziness and Soft Computing”, Springer, 2002, https://doi.org/10.1007/978-3-7908-1805-5

[14] G. Gursel, “Fuzzy Logic in Healthcare”, Medical Imaging: Concepts, Methodologies, Tools, and Applications, pp. 1-29, 2017, https://doi.org/10.4018/978-1-5225-0571-6.ch006

[15] K. B. Sundharakumar, S. Dhivya, S. Mohanavalli, R. Vinob Chander, “Cloud Based Fuzzy Healthcare System”, Procedia Computer Science, vol. 50, 2015, pp. 143-148.

[16] H. Choi, K. Han, K. Choi, and J. Y. Ahn, “A Fuzzy Medical Diagnosis Based on Quantiles of Diagnostic Measures”, Journal of Intelligent and Fuzzy Systems, vol. 31, no. 6, pp. 3197-3202, 2016.

[17] N. Allahverdi, S. Torun, and I. Saritas, “Design of a Fuzzy Expert System for Determination of Coronary Heart Disease Risk”, CompSysTech’07, pp. 1-8, 2007.

[18] S. Senthilnathan, “Usefulness of Correlation Analysis”, SSRN Electronic Journal, pp. 1-9, 2019.

[19] D. Oto-Peralías, “Regional Correlations of COVID-19 in Spain” [preprint], pp. 1-28, 2020, https://doi.org/10.31219/osf.io/tjdgw

[20] U. V. Mahajan, M. Larkins-Pettigrew, “Racial Demographics and COVID-19 Confirmed Cases and Deaths: A Correlational Analysis of 2886 US Counties”, Journal of Public Health, pp. 1-4, 2020, https://doi.org/10.1093/pubmed/fdaa070

[21] H. Eriksson et al., “Ontology Based Modeling of Pandemic Simulation Scenarios”, Studies in health technology and informatics, 129(Pt 1), pp. 755-759, 2007.

[22] C. Pesquita, J. Ferreira, F. Couto, and M. Silva, “The Epidemiology Ontology: An Ontology for the Semantic Annotation of Epidemiological Resources”, Journal of Biomedical Semantics, 2014, 5:4, pp. 1-7.

[23] S. M. Babcock, J. Beverley, L. G. Cowell, and B. Smith, “The Infectious Disease Ontology in the Age of COVID-19” [preprint], pp. 1-33, 2020, https://doi.org/10.31219/osf.io/az6u5

[24] N. Petrovic, M. Tosic, “Never-Ending Ontology Learning Approach Applied to Biomolecular Function Prediction”, IcETRAN 2019, pp. 762-767, 2019.

[25] N. Petrović, M. Tošić, “Explainable Artificial Intelligence and Reasoning in Smart Cities”, YuInfo 2020, pp. 1-6, 2020.

[26] H. H. Musa, N. Abdelrhman, “Comparing the Ranking Performance of Page Rank Algorithm and Weighted Page Rank Algorithm”, Journal of Computational and Theoretical Nanoscience, vol. 24, no. 1, pp. 750-753, 2018.

[27] 5G - Statistics & Facts [online]. Available on: https://www.statista.com/topics/3447/5g/

[28] N. Petrović, “Simulation Environment for Optimal Resource Planning During COVID-19 Crisis”, ICEST 2020.

[29] N. Petrović, M. Radenković, and V. Nejković, “Data-Driven Mobile Applications Based on AppSheet as Support in COVID-19 Crisis”, IcETRAN 2020, pp. 1-6, 2020.